UNIVERSIDADE FEDERAL DE PERNAMBUCO

CENTRO DE BIOCIÊNCIAS

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA VEGETAL

LUCAS ALEXANDRE DE SOUZA COSTA

DIVERSIFICAÇÃO GENÔMICA EM *RHYNCHOSPORA* VAHL. (CYPERACEAE), UM GÊNERO COM CROMOSSOMOS HOLOCÊNTRICOS

Recife

2022

LUCAS ALEXANDRE DE SOUZA COSTA

DIVERSIFICAÇÃO GENÔMICA EM *RHYNCHOSPORA* VAHL. (CYPERACEAE), UM GÊNERO COM CROMOSSOMOS HOLOCÊNTRICOS

Tese apresentada ao Programa de Pós-Graduação em Biologia Vegetal da Universidade Federal de Pernambuco, como um requisito parcial para a obtenção do título de doutor em Biologia Vegetal.

Área de concentração: Sistemática e Evolução.

Orientador: Prof. Dr. Luiz Gustavo Rodrigues Souza

(Depto. de Botânica, UFPE)
Co-orientadora: Profa. Dra. Andrea Pedrosa Harand

(Depto. de Botânica, UFPE)

Recife
2022

Catalogação na Fonte:
Bibliotecária Natália Nascimento, CRB4/1743

Costa, Lucas Alexandre de Souza.
Diversificação genômica em *Rhynchospora* VAHL. (cyperaceae), um gênero com cromossomos holocêntricos. / Lucas Alexandre de Souza Costa. – 2022.

187 f. : il., fig.; tab.

Orientador: Luiz Gustavo Rodrigues Souza.
Coorientadora: Andrea Pedrosa Harand.

Tese (doutorado) – Universidade Federal de Pernambuco. Centro de Biociências. Programa de Pós-graduação em Ciências Biológicas, Recife, 2022.
Inclui referências.

1. DNA repetitivo. 2. DNA satélite - elementos transponíveis. 3. Sequenciamento por captura de alvo. 4. Cromossomos holocêntricos. 5. Métodos comparativos filogenéticos. I. Souza, Luiz Gustavo Rodrigues. (orient.). II. Harand, Andrea Pedrosa. (coorient.). III. Título.

587 CDD (22.ed.) UFPE/CB – 2022-253

LUCAS ALEXANDRE DE SOUZA COSTA

DIVERSIFICAÇÃO GENÔMICA EM *RHYNCHOSPORA* VAHL. (CYPERACEAE), UM GÊNERO COM CROMOSSOMOS HOLOCÊNTRICOS

Tese apresentada ao Programa de Pós-Graduação em Biologia Vegetal da Universidade Federal de Pernambuco, como um requisito parcial para a obtenção do título de doutor em Biologia Vegetal.

Apresentada em: 24/02/2022

BANCA EXAMINADORA

Prof. Dr. Luiz Gustavo Rodrigues Souza (Orientador)
Universidade Federal de Pernambuco

Prof. Dr. André Luís Laforga Vanzela (Titular Externo)
Universidade Estadual de Londrina, Depto. de Biologia Geral

Prof. Dr. Diogo Cavalcanti Cabral de Mello (Titular Externo)
Universidade Estadual Paulista, Depto. de Biologia Geral e Aplicada

---

Prof[a]. Dr[a]. Giovana Augusta Torres (Titular Externo)
Universidade Federal de Lavras, Departamento de Biologia

Prof. Dr. Lyderson Facio Viccini (Titular Externo)
Universidade Federal de Juiz de Fora, Departamento de Biologia

Recife
2022

AGRADECIMENTOS

Ops, vou escrever demais...

Quatro anos depois de ter entregue minha dissertação de mestrado, eu já havia esquecido quão difícil é escrever essa sessão específica. Talvez eu deveria apenas listar todas as pessoas que me ajudaram a trilhar esse caminho, mas a minha veia de contador de história me impede de fazer algo mais simples. E hoje tem apenas uma história que eu queria contar. Uma história contada pelo meu pai. De como estava faltando comida em casa, de como as contas estavam todas vencidas, de como minha mãe passou dias na máquina de costura fazendo mochilas dos "bananas de pijama". De como meu pai precisou de toda a coragem que tinha pra ir tentar vender essas bolsas no centro de Maceió. De como isso virou a profissão dele, nosso sustento. De como os colegas camelôs dele o aconselhavam a trazer o filho pra ajudar nas vendas, e da resposta dele: "não, meu filho tem que estudar. Tem que ser doutor." Parte de mim queria ter ajudado, parte de mim, especialmente hoje, entende a nobreza no trabalho que eles faziam, o trabalho que nos carregou por anos. Vejo o quanto fui blindado, o quanto no meio de tanto aperto e sofrimento, fui privilegiado de certa forma. O mínimo que eu podia fazer é o que estou fazendo hoje, virando o doutor que meus pais fizeram de tudo para eu ser. E por isso que sou grato a eles mais do que tudo e todos! Por isso que essa tese é deles.

Muito se aprende e muito se muda em quatro anos. Especialmente quando metade deles são vividos em meio ao maior desafio da geração, a um período tão triste e sombrio como o da pandemia. É difícil encontrar motivação para seguir em frente quando todas as feridas do mundo são tão escancaradas por um evento, quando o negacionismo, a falta de empatia, a maldade, predominam tanto. Ser cientista, especialmente, parece um desafio maior do que jamais foi. E é por isso que agradeço aos meus colegas de laboratório, tanto os atuais como tantos que já passaram: Brena (minha madrinha preferida), Erton, Amália, Amandão, Amandinha, Cláudio, Thiago, Paulo, Yennifer, Natália, Jéssica, Gustavinho, Ana, Géssica, Bruna, Mariela, Pablo, e tantos outros! Vocês todos me inspiram, me motivam a persistir, a ser o melhor pesquisador que eu posso ser! Muito obrigado!

Eu demorei muito a entender que antes dos outros acreditarem em mim, eu deveria acreditar primeiro. Porém, ter orientadores como Gustavo e Andrea foi essencial para que eu aprendesse a confiar no meu trabalho. Vocês não fazem ideia de como a sua confiança em mim e na minha capacidade me ajudaram a ser o pesquisador que sou hoje. Não sabem quantas vezes eu pensei em desistir, pensei que não era capaz. E sabe o que me tirava esses pensamentos?

Bom, terapia. E também conversar com vocês dois. Vocês sempre me ajudaram a encontrar a força, motivação e confiança necessárias para continuar. O que eu sou hoje e o que eu venha a me tornar, devo a vocês dois!

A pesquisa nunca é feita por apenas uma pessoa. Colaboração é essencial, é chave. Uma das coisas que mais me orgulho de todos os meus anos de pós-graduação é a quantidade de parcerias que estabeleci, a quantidade de artigos e projetos diferentes dos quais pude participar. A pesquisa me fez viajar, literalmente e figurativamente também. Passar oito meses na Alemanha sobre a tutela de uma lenda da área, Dr. Andreas Houben (meu alemão preferido), com certeza moldou não só meu caráter profissional como vários aspectos pessoais também. Sou grato a todos os meus parceiros no Brasil e fora, pela confiança no meu trabalho e por tudo que aprendi e aprendo com vocês.

Bom, por mais que as pressões externas e internas tentem nos fazer acreditar, nem só de ciência vive o pesquisador. O período de pandemia me deixou longe de vários amigos, mas me aproximou de vários outros. Nunca imaginei que na minha idade estaria fazendo amiguinhos em jogos on-line, mas hoje tem uma lista de pessoas que eu nunca vi pessoalmente, mas que falo todos os dias! E além disso, um velho amigo que a internet me ajudou a reaproximar! Pessoas que não só aliviam o meu estresse diário com a leveza das nossas noites no Discord e piadas no Whatsapp, mas que também me motivam e torcem tanto por mim! Obrigado, meus *Warriors of light*!

Eu disse que mudei muito nos últimos quatro anos né? Eu disse que eu casei? Eu disse que eu casei com o amor da minha vida? Por quê isso aconteceu também. Por 10 anos, Yhanndra foi não só uma namorada, ela foi uma colega, uma inspiração, minha confidente e melhor amiga. Estou compartilhando com ela o sonho de me tornar um doutor, mas antes disso realizei o maior sonho de todos, que era casar com ela. Nosso sonho do nosso cantinho, de não precisar dizer “tchau” no fim do dia. Independente do que acontecer daqui pra frente, esse sonho nós já realizamos. Te amo pra sempre!

Bom, uma tese não se faz só com amizades e orientações, se faz com muita burocracia e com dinheiro. Então fica também meu agradecimento ao PPGBV, à UFPE e aos órgãos de fomento, que mesmo sob o comando do PIOR governo da história desse país, seguem na luta pela ciência. E a ciência há de resistir, mesmo quando todo o resto cair.

”And there may come a day when you forget the faces
and voices of those you have met along the way.
On that day, I bid you remember this:
That no matter how far your journey may take you,
you stand where you stand by virtue of the
road you walked to get there.
For in times of hardship, when you fear you cannot go on…
The joy you have known, the pain you have felt,
the prayers you have whispered and answered-
-they shall ever be your strength and your comfort.”

Gra’ha Tia, Final Fantasy XIV

RESUMO

Com os avanços nas técnicas de sequenciamento genômico, foi descoberto que a maior parte do genoma das plantas é composto por DNA repetitivo. Dentre esta fração do genoma, se destacam os DNAs Satélite (longos arranjos de unidades de repetição chamadas monômeros) e os elementos transponíveis (DNA repetitivo disperso com a habilidade de criar cópias e se "mover" ao longo do genoma hospedeiro). Apesar de não codificarem proteínas essenciais para o hospedeiro, tem sido visto que tanto DNAs satélites como elementos transponíveis podem impactar a evolução do genoma hospedeiro, alterando o espaço genético/epigenético e promovendo funções estruturais para determinadas regiões cromossômicas. Além disso, a variabilidade na abundância de diferentes tipos de DNA repetitivo pode estar ligada a respostas naturais destes a estresses ambientais. Nesta tese, buscamos investigar o impacto de diferentes DNAs repetitivos na evolução genômica e diversificação de *Rhynchospora* Vahl (Cyperaceae), um gênero bastante diverso (~400 esp.) com ampla distribuição geográfica. Espécies de *Rhynchospora* se destacam por possuírem cromossomos holocêntricos (centrômero disperso ao longo de toda a extensão das cromátides), apresentando nestes o DNA satélite *Tyba*, primeiro satélite específico do centrômero a ser descoberto em uma espécie holocêntrica. No primeiro capítulo, investigamos a possibilidade de utilizar o subproduto de sequenciamento por captura de alvo para a identificação, localização cromossômica e filogenômica comparativa de DNA repetitivo. Nós validamos os resultados mediante comparação dos dados obtidos por captura de alvo com dados de *genome skimming*, mais tradicionalmente utilizados para o estudo de sequências repetitivas. No segundo capítulo, nós estudamos a evolução do satélite holocentromérico *Tyba* em uma ampla amostragem do gênero. Nossos resultados mostram um alto sinal filogenético para *Tyba* em *Rhynchospora* e alta conservação, provavelmente ligada à uma possível vantagem estrutural promovida aos holocentrômeros. No terceiro capítulo, utilizamos métodos comparativos filogenéticos a fim de investigar diferentes fatores que possam ter influenciado a abundância de retroelementos LTR nos genomas de *Rhynchospora*. Foi visto que a abundância desses elementos é impactada por uma combinação de fatores genômicos, temporais, ambientais e, principalmente, filogenéticos. No geral, nossos resultados contribuem para um maior conhecimento do impacto de *Tyba* e de retroelementos LTR na evolução genômica de *Rhynchospora*, além de demonstrar o potencial do estudo de elementos repetitivos num contexto macroevolutivo.

Palavras-chave: DNA Repetitivo. DNA satélite. Elementos Transponíveis. Sequenciamento por captura de alvo. Cromossomos Holocêntricos. Métodos Comparativos Filogenéticos.

ABSTRACT

With the advent of Next Generation Sequencing, it was shown that most of plants genomes are composed by repetitive DNA. Within this genomic fraction, two specific types stand out: Satellite DNAs (long arrays of tandemly arranged repetitive units known as monomers) and transposable elements (dispersed repetitive DNA with possessing the ability to create copies and "move" along the host genome). Although these sequences do not code essential proteins for the host, it has been shown that satellite DNA and transposable elements can impact genome evolution, altering the genetic/epigenetic landscape and promoting structural functions for specific chromosome regions. In addition to this, the variability in the abundance of different repetitive DNAs can be linked to its natural response to environmental stress. In this thesis, we aim to investigate the impact of different repetitive sequences on the genome evolution and diversification of *Rhynchospora* Vahl., a highly diverse (~400 spp.) and widely distributed genus. Species of *Rhynchospora* stand out for presenting holocentric chromosomes (with the centromere disperse along all the extension of the chromosomes), with the presence of satellite DNA *Tyba*, the first centoomere-specific satellite discovered in a holocentric species. In the first chapter, we investigate the possibility of utilizing the byproduct of target capture sequencing for the identification, chromosomal mapping and comparative phylogenomics of repetitive DNA. We validate this approach by comparing the results obtained by target capture with data acquired from methods more traditionally used for repetitive sequence study. In the second chapter, we studied the evolution of holocentromere satellite DNA *Tyba* in a large sampling of the genus. Our results showed a high phylogenetic signal for *Tyba* in *Rhynchospora* and high sequence conservation, probably related to structural advantages provided by the satellite to the holocentromeres. In the third chapter, we used phylogenetic comparative methods to investigate different factors that may have influenced the abundance of LTR-retroelements in the *Rhynchospora* genomes. We showed that the abundance of these elements is impacted by a combination of genomic, temporal, environmental and, most importantly, phylogenetic factors. In general, our results have contributed to a broader understanding of the impact of *Tyba* and LTR-retroelements in *Rhynchospora*, while also demonstrating the potential of studying repetitive DNA in a macroevolutionary context.

Keywords: Repetitive DNA. Satellite DNA. Transposable Elements. Target-capture Sequencing. Holocentric Chromosomes. Phylogenetic Comparative Methods

SUMÁRIO

1 INTRODUÇÃO 12

2 FUNDAMENTAÇÃO TEÓRICA 15

2.1 SEQUENCIAMENTO GENÔMICO 15

2.1.1 Introdução 15

2.1.2 Sequenciamento de Nova Geração 16

2.2 DIVERSIFICAÇÃO GENÔMICA 18

2.2.1 Número e morfologia cromossômica 18

2.2.2 Tamanho do genoma 20

2.2.3 DNA Repetitivo 21

2.2.4 DNA repetitivo em tandem 22

2.2.5 DNA repetitivo disperso 24

2.3 CROMOSSOMOS HOLOCÊNTRICOS 25

2.3.1 Estrutura e organização 25

2.3.2 DNA repetitivo centromérico 27

2.3.3 Cromossomos holocêntricos em *Rhynchospora* Vahl. (Cyperaceae) 28

2.4 MÉTODOS COMPARATIVOS FILOGENÉTICOS 30

2.4.1 Introdução 30

2.4.2 Taxa de diversificação 31

2.4.3 Análises de correlação 32

2.4.4 Filogenômica Comparativa 33

3 RESULTADOS 35

3.1 WHAT DETERMINES RATES OF SATDNA EVOLUTION? INVESTIGATING THE DIVERSIFICATION OF HOLOCENTROMERIC SATDNA *TYBA* IN *RHYNCHOSPORA* (CYPERACEAE) 35

3.2 WHAT DRIVES LTR-RETROTRANSPOSON EVOLUTION IN *RHYNCHOSPORA* (CYPERACEAE) GENOMES? 90

4 CONCLUSÕES 130

REFERÊNCIAS 131

APÊNDICE A – AIMING OFF THE TARGET: RECYCLING TARGET CAPTURE SEQUENCING READS FOR INVESTIGATING REPETITIVE DNA 146

1 INTRODUÇÃO

Desde a descoberta do papel essencial do DNA para a hereditariedade, a composição do genoma tem sido tópico crucial de diversas pesquisas científicas (COBB, 2014). O sequenciamento genômico de nova geração revolucionou a forma como compreendemos o genoma dos eucariotos (SUN et al., 2021). Além de várias metodologias permitirem o sequenciamento de genomas completos, uma grande diversidade de métodos de "complexidade reduzida" tem sido essencial para a aquisição rápida e de baixo custo de um grande número de sequências (ALLEN et al., 2017; LEMOPOULOS et al., 2019; MANDEL et al., 2014; MARGUERAT; BÄHLER, 2010; STRAUB et al., 2012). Estas sequências podem ser usadas tanto para fins sistemáticos quanto para a caracterização de frações genômicas de interesse, como o DNA repetitivo (DODSWORTH, 2015; DODSWORTH et al., 2019). Graças aos avanços nas técnicas de sequenciamento, hoje sabemos que a maior parte do genoma dos eucariotos, em especial das plantas, é composto por DNA repetitivo (ELLIOTT; GREGORY, 2015). Tradicionalmente, dois tipos de DNA repetitivo são amplamente estudados em plantas: DNAs satélites, sequências repetidas em tandem formada por longos arranjos de unidades de repetição conhecidas como monômeros (GARRIDO-RAMOS, 2017; PLOHL; MEšTROVIC; MRAVINAC, 2012) e elementos transponíveis, sequências dispersas no genoma que recrutam a maquinaria enzimática celular para produzir cópias e se proliferar no genoma hospedeiro (ARKHIPOVA, 2017; PIÉGU et al., 2015). Apesar de, por muitos anos, os diversos tipos de DNA repetitivo serem considerados "DNA lixo" e sequências "egoístas" que apenas contribuíam para o aumento acelerado do tamanho do genoma (EDDY, 2012; GEMMELL, 2021), estudos recentes têm mostrado o potencial destes na evolução e diversificação dos genomas. Por exemplo, é comum encontrar DNAs satélites e/ou elementos transponíveis específicos à regiões cromossômicas funcionais como os centrômeros, levantando a hipótese que estas sequências promovam alguma vantagem estrutural à estas regiões (ÁVILA ROBLEDILLO et al., 2018; CHENG et al., 2002; MARQUES et al., 2015; NAGAKI et al., 2003; PLOHL; MEŠTROVIĆ; MRAVINAC, 2014). Além disso, tem sido mostrado que a mobilidade dos elementos transponíveis efetivamente afeta o espaço gênico e epigenético, promovendo mudanças em genes (BENNETZEN; WANG, 2014; SCHRADER; SCHMITZ, 2019). Além disso, diversos estudos têm relacionado variações na abundancia de diferentes DNAs repetitivos com distribuição geográfica e nicho ecológico (BILINSKI et al., 2017; DÍEZ et al., 2013; LYU et al., 2018; SCHLEY et al., 2021).

Neste contexto, o gênero *Rhynchospora* Vahl. (Cyperaceae) aparece como um interessante modelo para o estudo da evolução e dinâmica do DNA repetitivo. *Rhynchospora* é composto por aproximadamente 400 espécies com uma ampla distribuição geográfica, com origem estimada na América do Sul e posterior migração e diversificação no hemisfério norte (BUDDENHAGEN, 2016; SPALINK et al., 2016; THOMAS; ARAÚJO; ALVES, 2009). Cariotipicamente, *Rhynchospora* se destaca pela presença de cromossomos holocêntricos, um tipo de cromossomo que possui o centrômero disperso ao longo das cromátides ao invés de localizado em uma região específica como na maioria dos eucariotos (VANZELA; GUERRA, 2000). Em um estudo citogenômico, Marques et al. (2015) descobriram em *Rhynchospora pubera* o primeiro DNA satélite específico ao centrômero de uma espécie holocêntrica, denominado *Tyba*. Estudos posteriores mostraram que *Tyba* está presente em outras espécies do gênero como *R. tenuis* e *R. ciliata*, mas não aparece em todo o gênero (RIBEIRO et al., 2017). Além disso, foi mostrado que eventos cromosssômicos como poliploidias e disploidias moldaram a evolução cromossômica do grupo, além de promover uma variabilidade moderada de tamanhos de genoma (BURCHARDT et al., 2020; RIBEIRO et al., 2018). Apesar dessa interessante diversidade genômica, somada à ampla diversidade e distribuição geográfica do grupo tornam *Rhynchospora* um grupo interessante para o estudo de DNA repetitivo em um contexto macroevolutivo, poucas espécies do grupo possuem caracterização da fração repetitiva disponível. Recentemente, usando uma abordagem de sequenciamento por captura de alvo, Buddenhagen (2016) gerou sequências de mais de 100 espécies de *Rhynchospora*. Apesar dessa metodologia ser direcionada para regiões-alvo específicas de cópia única, foi possível utilizar o subproduto deste sequenciamento para a obtenção de sequências plastidiais e DNA ribossômico, além de ser especulado que esse subproduto sirva para caracterizar sequências altamente repetitivas (DODSWORTH et al., 2019).

Nesta tese buscamos estudar a diversificação genômica no gênero *Rhynchospora* com ênfase em sua fração repetitiva. Para isto, a tese encontra-se estruturada em três capítulos. No primeiro capítulo, validamos o uso do subproduto de sequenciamento por captura de alvo para o estudo de DNA repetitivo, focando na caracterização e identificação de sequências, produção de sondas para estudos citogenômicos e análises filogenômicas baseadas em sequências repetitivas. No segundo capítulo, estudamos a evolução de *Tyba*, DNA satélite específico aos holocêntromeros de *Rhynchospora*, em uma ampla amostragem do grupo, encontrando uma alta conservação de sequência para este satélite e discutindo possíveis fatores relacionados à esta conservação. No terceiro capítulo, procuramos entender se fatores como a abundância de

Tyba, tempo de diversificação, variáveis ecológicas e relações filogenéticas impactam a abundância de elementos transponíveis (particularmente retrotransposons LTR) dentro do gênero utilizando métodos comparativos filogenéticos.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 SEQUENCIAMENTO GENÔMICO

2.1.1 Introdução

O ácido desoxirrribonucleico (DNA, do inglês *desoxyribonucleic acid*) é uma das mais importantes moléculas para a vida na Terra, pois carrega em sua composição as informações necessárias para a hereditariedade (HEATHER; CHAIN, 2016). Considerado por muitos o pai do sequenciamento, Frederick Sanger, em seus estudos no fim da década de 1940 e início da década de 1950, mostrou a importância do conceito de "sequência" para a biologia, expondo o fato de que as proteínas eram compostas por polipeptídeos formados por resíduos de aminoácidos arranjados em uma ordem predefinida (SANGER, 1949; SANGER; TUPPY, 1951). Pouco após, a estrutura tridimensional do DNA foi resolvida no emblemático estudo de Watson e Crick no início da década de 1950 (WATSON; CRICK, 1953) com contribuição fundamental dos achados de Rosalin Frankilin e Maurice Wilkins (ZALLEN, 2003).

Apesar desses avanços, 15 anos foram decorridos desde a descoberta da estrutura de dupla hélice do DNA até o primeiro sequenciamento bem-sucedido (KAISER; WU, 1968; WU; KAISER, 1968). Vários fatores limitantes levaram a este amplo intervalo, tais como a similaridade bioquímica dos constituintes do DNA e a dificuldade de separá-los, o tamanho excepcional da molécula quando comparada ao tamanho de proteínas e a dificuldade para quebrar as moléculas de DNA sem o conhecimento, na época, de DNAses específicas (HUTCHISON, 2007). O sequenciamento "Sanger", batizado com o nome de um de seus idealizadores, contornou alguns destes problemas com uma técnica onde fragmentos sintetizados de DNA eram separados por eletroforese em gel de poliacrilamida, aumentando assim a cadeia de reação e tamanho das moléculas sequenciadas (SANGER; COULSON, 1975). Posteriormente, o uso de diferentes fluorocromos que se ligavam aos nucleotídeos finais dos fragmentos de sequência permitiu a automação da leitura das sequências de DNA separadas nos géis de poliacrilamida (SMITH et al., 1986). A substituição dos géis de poliacrilamida por capilares para as reações de eletroforese aumentou mais ainda a rapidez e o rendimento do sequenciamento Sanger, automatizando-o por completo (HUTCHISON, 2007). Essa "primeira geração" de métodos de sequenciamento permitiu importantes marcos na história da genética,

como o primeiro genoma animal completamente sequenciado [*Caenorhabditis elegans* (THE C. ELEGANS SEQUENCING CONSORTIUM, 1998)] e o sequenciamento completo do genoma humano (INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM et al., 2001; VENTER et al., 2001).

2.1.2 Sequenciamento de Nova Geração

Mesmo com todos os avanços do método Sanger, os sequenciamentos de DNA de "primeira geração" ainda eram constituídos de um processo lento e trabalhoso. Essa necessidade de metodologias de sequenciamento com um rendimento ainda maior levou ao surgimento de técnicas que buscavam analisar um grande número de sequências em paralelo, dando origem aos métodos de "sequenciamento de nova geração" (NGS, do inglês *Next Generation Sequencing*) (VAN DIJK et al., 2014). O primeiro método a cruzar esta barreira geracional foi o chamado "pirosequenciamento", usado pelo sequenciador automático 454 desenvolvido pela 454 Life Sciences (MARGULIES et al., 2005). Esta técnica permite a análise simultânea de milhares de fragmentos de DNA capturados em pequenas esferas (beads), baseando-se na quantidade de pirofosfato liberada por cada nova base de DNA sintetizada (RONAGHI et al., 1996). A quantidade massiva de sequências analisadas paralelamente pelo método do 454 foi superada pelos sequenciadores Illumina, desenvolvidos pela Solexa/Illumina (VOELKERDING; DAMES; DURTSCHI, 2009). Nestes aparelhos, moléculas de DNA ligadas a adaptadores são fixadas em uma placa e amplificadas por PCR, formando *clusters* que são utilizados como modelos por DNA polimerases que adicionam bases complementares em ciclos, liberando fluróforos que são detectados e lidos por sensores (VOELKERDING; DAMES; DURTSCHI, 2009). Apesar dos sequenciadores 454 e Illumina dominarem essa primeira onda de equipamentos de nova geração (também chamadas de métodos de segunda geração), outras plataformas como a SOLiD (do inglês *sequencing by oligonucleotide ligation and detection*) e Ion Torrent apareceram como alternativas competitivas (HEATHER; CHAIN, 2016).

Uma nova mudança de paradigma no que diz respeito à técnicas de sequenciamento aconteceu com o surgimento das chamadas metodologias de "terceira geração", geralmente caracterizadas pela capacidade de sequenciamento de moléculas únicas, sem a necessidade de amplificação de DNA (HEATHER; CHAIN, 2016). A primeira metodologia de sequenciamento de molécula única foi comercializada pela Helicos BioSciences, que utilizava

uma técnica de fixação da molécula de DNA à uma superfície similar à Illumina, mas lendo uma base por ciclo, sem a necessidade de amplificação (BRASLAVSKY et al., 2003). Outro grande avanço foi o advento do sequenciamento em tempo real de molécula única (SMRT, do inglês *single molecule real time*) promovido pelos aparelhos PacBio da Pacific Biosciences (VAN DIJK et al., 2014). Esses aparelhos permitem o sequenciamento em tempo real de longas moléculas únicas de DNA por meio da leitura de comprimentos de onda emitidos pela síntese de novos nucleotídeos complementares (LEVENE et al., 2003). Por fim, nos últimos anos, a aplicação de princípios de "nanoporos" ao sequenciamento apareceu como uma possível revolução na genômica, permitindo o sequenciamento de grandes moléculas de DNA que passam por uma "nano-abertura", emitindo voltagens nucleotídeo-específicas (HAQUE et al., 2013). O sequenciamento por nanoporo tem garantido não só altíssima resolução e rapidez, como também mobilidade ao processo de sequenciamento, com a disponibilidade de sequenciadores portáteis no mercado (JAIN et al., 2016).

2.1.2.1 Métodos de complexidade reduzida

No geral, as tecnologias de segunda e terceira geração revolucionaram o sequenciamento genômico, promovendo um aumento exponencial no número de genomas completos sequenciados (SUN et al., 2021). Os impactos destes avanços tecnológicos estão presentes não só na área da genética/genômica, mas também em estudos de biodiversidade, como a filogenia, sistemática e ecologia (DODSWORTH, 2015). Enquanto que genomas completos são essenciais para uma maior compreensão da evolução de características e detalhamento de diferenças genéticas entre espécies, seu uso para estudos sistemáticos e filogenéticos ainda é limitado, sendo a complexidade genômica um importante fator limitante (DODSWORTH et al., 2019). Dessa forma, diferentes metodologias têm focado na redução dessa complexidade, de forma a diminuir a quantidade de dados gerados e o custo envolvido, além de facilitar a análise bioinformática desses dados para fins sistemáticos (ALLEN et al., 2017).

Uma das metodologias de complexidade reduzida mais utilizadas é o chamado "*genome skimming*", um tipo de sequenciamento superficial onde é definida uma cobertura baixa de sequenciamento em relação ao total de DNA de uma espécie (DODSWORTH, 2015). Essa baixa cobertura (0.1 até 0.4x do genoma) é normalmente suficiente para a caracterização da fração de DNA repetitivo de uma espécie, além de também ser suficiente para a montagem de plastomas e DNAs ribossomais (STRAUB et al., 2012). Outra metodologia que vem ganhando

espaço em estudos sistemáticos é o sequenciamento de transcriptoma, também conhecido como RNA-seq (MARGUERAT; BÄHLER, 2010). O RNA-seq tem como objetivo o sequenciamento de DNA complementar, permitindo não apenas o estudo detalhado da expressão gênica de um organismo como também o uso de um grande conjunto de genes para a construção de filogenias (WANG et al., 2017). Por outro lado, o sequenciamento associado a sítio de restrição (RAD-seq, do inglês *Restriction site associated DNA sequencing*) surge como uma alternativa inteiramente voltada à sistemática, especialmente em grupos com diversificação recente (LEMOPOULOS et al., 2019). O RAD-Seq inova ao utilizar enzimas de restrição para cortar e selecionar pequenas regiões de DNA adjacentes a sítios de restrição, com posterior amplificação e sequenciamento dessas regiões (DAVEY et al., 2011).

Nos últimos anos uma metodologia específica tem ganhado cada vez mais atenção entre os métodos de complexidade reduzida: o sequenciamento por captura de alvo (ANDERMANN et al., 2020). Este método se baseia no desenvolvimento de sondas de DNA que são usadas para capturar regiões de cópia única complementares específicas (regiões alvo) de um determinado genoma (ALBERT et al., 2007; GNIRKE et al., 2009). Graças à rapidez do método e à alta quantidade de dados gerados, chegando à centenas de regiões sequenciadas por espécie, métodos de captura de alvo têm sido amplamente utilizados em estudos filogenéticos (HEYDUK et al., 2016; MANDEL et al., 2014; OGUTCEN et al., 2021; SINISCALCHI et al., 2019). Um dos maiores atrativos desse sequenciamento é a possibilidade do desenvolvimento de conjuntos universais de sondas, isto é, sondas que possam servir para sequenciar várias ou todas as espécies de um grupo (BUDDENHAGEN et al., 2016; CHAFIN; DOUGLAS; DOUGLAS, 2018; JOHNSON et al., 2019). Outro ponto positivo do sequenciamento por captura de alvo é a possibilidade da utilização de sequências não-alvo que acabam sendo capturadas no processo, técnica comumente chamada de sequenciamento híbrido (Hyb-seq, do inglês *Hybrid Sequencing*) (WEITEMIER et al., 2014). Essa fração não-alvo é comumente similar a um sequenciamento por *genome skimming*, abrindo a possibilidade de que o sequenciamento híbrido sirva tanto para a captura de regiões gênicas como para a caracterização de outras sequências, tais como DNA plastidial, DNA ribossomal e, potencialmente, DNA repetitivo (DODSWORTH et al., 2019; SCHMICKL et al., 2016; SPROUL; BARTON; MADDISON, 2020).

2.2 DIVERSIFICAÇÃO GENÔMICA

2.2.1 Número e morfologia cromossômica

O cromossomo é o último nível de condensação da cromatina, estrutura formada pelo DNA e proteínas presentes no núcleo eucarioto. O cariótipo (conjunto de cromossomos de um organismo) ganha destaque em estudos evolutivos por ser uma representação direta do próprio genoma (GUERRA, 2012). Somado à alta diversidade morfológica e de nicho ecológico, as angiospermas representam um dos grupos com maior variabilidade cromossômica, apresentando um histórico evolutivo marcado por disploidias (perda ou ganho de um cromossomo por eventos de fissão/fusão) e poliploidia (duplicação do complemento cromossômico) (GLICK; MAYROSE, 2014; GUERRA, 2016; JIAO et al., 2011). Desta forma, desde os trabalhos pioneiros de Avdulov (1931) e Stebbins (1966), a relação entre variabilidade cromossômica, plasticidade ecológica e eventos de especiação têm sido um dos focos da citogenética vegetal. Neste contexto de diversificação, eventos de poliploidia se destacam pela rápida geração de plasticidade genômica, levando ao aparecimento de novidades evolutivas que levam a especiação e colonização de novos ambientes (ALIX et al., 2017; BURGESS et al., 2014; ĆERTNER et al., 2017; JIAO et al., 2011). Não é por menos que o próprio surgimento da flor tem sido associado a um evento de poliploidia na base da diversificação das angiospermas (DODSWORTH; CHASE; LEITCH, 2016). Diversos outros estudos têm demonstrado a correlação de eventos de poliploidia com aumentos na taxa de diversificação em larga escala das angiospermas como um todo (LANDIS et al., 2018; TANK et al., 2015) e em casos pontuais, como no gênero *Passiflora* (SADER et al., 2019).

Além da aparente correlação entre poliploidia e diversificação, relações entre variação numérica e latitude já foram relatadas em diversos grupos, mostrando um possível valor adaptativo do número cromossômico na colonização de novos habitats (BEDINI; GARBARI; PERUZZI, 2012; PERUZZI et al., 2012). A questão, no entanto, parece não estar apenas relacionada ao número cromossômico, mas também à sua morfologia. Avdulov (1931) e Stebbins (1966) já relatavam que cromossomos de espécies de regiões tropicais eram significativamente menores que os de espécies de regiões temperadas. Desde então, diversos trabalhos têm apresentado dados confirmando (LEVIN, 2012; SOUZA et al., 2019) ou contrariando (RAYBURN et al., 1985) a hipótese de Stebbins. Outro aspecto morfológico interessante do ponto de vista evolutivo é a organização do centrômero, que em eucariotos pode formar uma constrição primária no cromossomo (monocêntrico) ou se apresentar disperso ao longo das cromátides (holocêntrico) (GUERRA et al., 2010). Cromossomos holocêntricos apresentam maior susceptibilidade a eventos de fissão cromossômica graças a herança diferenciada de fragmentos cromossômicos, levando à uma alta variabilidade de números

cromossômicos (ESCUDERO et al., 2015). Esta capacidade de gerar variabilidade cariotípica tem impulsionado estudos relacionando o aparecimento de cromossomos holocêntricos com aumento de diversificação e colonização de novos habitats (MÁRQUEZ-CORRO; ESCUDERO; LUCEÑO, 2018; ZEDEK; BUREŠ, 2018).

2.2.2 Tamanho do genoma

Antes mesmo dos revolucionários estudos de Watson, Crick, Rosalin e Wilkins sobre a estrutura da molécula de DNA (WATSON; CRICK, 1953; ZALLEN, 2003), pesquisadores já mostravam interesse pela quantidade de material genético presente nos genomas das espécies (SWIFT, 1950). Graças ao advento de técnicas de quantificação de conteúdo nuclear rápidas e precisas como a citometria de fluxo, estimativas mais recentes mostram dados para aproximadamente 15.000 espécies de eucariotos, dos quais aproximadamente 7542 são de angiospermas (GARCIA et al., 2014). Estas se destacam dentre os eucariotos por apresentarem uma variação de aproximadamente 2.400 vezes entre o menor (*Genlisea margaretae*, 1C = 0.06 pg) (GREILHUBER et al., 2006) e maior genoma reportado (*Paris japônica*, 1C = 152.23 pg) (PELLICER; FAY; LEITCH, 2010). Essa variabilidade se torna ainda mais interessante pelo fato de que, até o presente momento, pouco menos de 3% do total estimado de espécies de plantas tiveram seu genoma quantificado (PELLICER et al., 2018). Além dessa alta variação, o tamanho do genoma é uma característica fortemente correlacionada com traços potencialmente adaptativos como ciclo de vida, desenvolvimento e taxa de evolução molecular (BROMHAM et al., 2015). Sendo assim, não é surpresa que o interesse pelo potencial adaptativo do tamanho do genoma em angiospermas tenha crescido ao longo dos anos.

Em termos teóricos, tanto o aumento como a diminuição do tamanho do genoma podem levar a mudanças na taxa de especiação de um determinado grupo, a depender de uma série de fatores (KRAAIJEVELD, 2010; VINOGRADOV, 2003). Estudos recentes têm demonstrado que a diversidade de tamanhos do genoma está correlacionada com a especiação (PUTTICK; CLARK; DONOGHUE, 2015). Por conta desta relação, a investigação de fatores ecológicos possivelmente relacionados ao tamanho do genoma em um contexto evolutivo têm sido objetivo de investigação. Inspirados pelas observações de Avdulov (1931) e Stebbins (1966), diversos trabalhos já exploraram a correlação entre tamanho do genoma e latitude, seja esta positiva (KANG et al., 2014; SOUZA et al., 2019) ou negativa (GROTKOPP et al., 2004; SCHMUTHS, 2004). Correlações (geralmente negativas) com o

tamanho do genoma também foram observadas em clines de altitude, inclusive a um nível intraespecífico (DÍEZ et al., 2013; ZAITLIN; PIERCE, 2010). É importante notar que espécies distribuídas em diferentes altitudes/latitudes estão sujeitas à diferentes pressões ambientais, mostrando que fatores ecológicos podem estar impactando a variabilidade no tamanho do genoma observada nestes casos (CACHO et al., 2021; ENKE; FUCHS; GEMEINHOLZER, 2011; SCHLEY et al., 2021; SOUZA et al., 2019). Neste contexto, a investigação das sequências genômicas responsáveis por esta variabilidade têm sido fundamentais para uma compreensão integrativa de seu papel evolutivo.

2.2.3 DNA Repetitivo

Por muitos anos, um dos maiores questionamentos em relação à variabilidade de tamanho do genoma era a ausência de correlação entre o conteúdo de DNA e a complexidade dos organismos, aspecto conhecido como “Paradoxo do Valor-C” (EDDY, 2012; THOMAS, 1971). Estudos mais detalhados permitiram uma resolução parcial deste paradoxo, com a descoberta de que grande parte desta variação estava relacionada à presença de sequências de DNA (em sua maioria não-codificante) altamente repetitivas (BRITTEN; KOHNE, 1968; ELLIOTT; GREGORY, 2015). Apesar de inicialmente chamado de “DNA lixo”, estudos recentes comprovam que a ação do DNA repetitivo no genoma hospedeiro pode estar ligada a processos adaptativos, tais como mudanças no espaço gênico (ativando ou desativando genes), alterações epigenéticas e funções estruturais nos cromossomos (ACHREM; SZUĆKO; KALINKA, 2020; DODSWORTH; LEITCH; LEITCH, 2015; SCHRADER; SCHMITZ, 2019).

Os avanços nas técnicas de sequenciamento de nova geração têm permitido uma visão mais detalhada do DNA repetitivo e seu possível papel na diversificação dos organismos (ELLIOTT; GREGORY, 2015; GEMMELL, 2021). Uma importante característica do DNA repetitivo é a sua marcante diversidade de tamanho, composição e organização de sequências. Nesse contexto, plataformas como o RepeatExplorer (NEUMANN et al., 2019; NOVAK et al., 2013) se fazem essenciais, permitindo a caracterização e comparação de sequências repetitivas mediante sequenciamentos de baixa cobertura. Apesar desses métodos conferirem uma caracterização bastante específica, tradicionalmente podemos dividir os elementos repetitivos de acordo com sua organização no genoma em dois tipos: DNA repetitivo em tandem e DNA repetitivo disperso (WEISS-SCHNEEWEISS et al., 2015).

2.2.4 DNA repetitivo em tandem

DNAs repetitivos em tandem são caracterizados por apresentarem unidades de repetição de dezenas a centenas de pares de base (monômeros) localizadas adjacentes umas às outras (GARRIDO-RAMOS, 2017; LIM et al., 2013). Dentre estes, estão os DNA satélites, normalmente classificados de acordo com o tamanho de seus monômeros (RICHARD; KERREST; DUJON, 2008). A similaridade entre as sequências dos monômeros de diferentes DNAs satélite também costumam ser usadas para fins classificativos, a fim de agrupá-los em diferentes famílias/sub-famílias (GARRIDO-RAMOS, 2017). DNAs satélite representam um dos grupos de sequências mais dinâmicas do genoma por apresentarem uma alta taxa de evolução e rápidas mudanças tanto em sua estrutura como em sua abundância (LOWER et al., 2017; MACAS; MESZAROS; NOUZOVA, 2002).

Ao longo dos anos, várias teorias foram propostas para explicar a evolução dinâmica dos DNAs satélites. Uma dessas teorias propõe que sequências de DNA satélite tendem a acumular mutações que com o tempo são homogenizadas e fixadas na maioria dos monômeros do arranjo, em um processo de evolução em concerto (GARRIDO-RAMOS, 2015; PLOHL; MEšTROVIC; MRAVINAC, 2012). Esse mecanismo acaba por levar à diferenciação entre o "satelitoma" (conjunto de DNAs satélites) mesmo entre espécies evolutivamente próximas (AHMAD et al., 2020; BELYAYEV et al., 2019). Outra importante teoria para explicar a evolução dos DNAs satélites é o chamado "modelo da biblioteca", onde espécies próximas compartilham uma coleção ou "biblioteca" de diferentes famílias de DNAs satélite (FRY; SALSER, 1977; SALSER et al., 1976). Com o tempo, nas diferentes espécies, diferentes famílias de DNA satélite são amplificadas e se tornam dominantes naqueles genomas, gerando uma alta diversidade de satelitomas entre espécies próximas (GARRIDO-RAMOS, 2015; PLOHL; MEšTROVIC; MRAVINAC, 2012).

No geral, a alta variabilidade estrutural e de abundância compromete a utilização de DNAs satélites em um contexto macroevolutivo, visto que diversas famílias são espécie- ou gênero-específicas (MACAS et al., 2015; RICHARD; KERREST; DUJON, 2008). Porém, ainda que pouco comum, existem casos de DNAs satélites com sequências altamente conservadas sendo compartilhados por espécies dentro de um mesmo gênero, família e até mesmo ordens (MRAVINAC; PLOHL; UGARKOVIĆ, 2005; PETRACCIOLI et al., 2015; ROBLES et al., 2004). Além das implicações evolutivas, o estudo do conjunto de diferentes linhagens de DNA satélite encontrados em um grupo de organismos têm promovido uma maior compreensão sobre a organização do genoma (RUIZ-RUANO et al., 2016). Com os avanços na citogenômica,

DNAs satélites linhagem-específico têm sido amplamente utilizados como eficientes marcas citomoleculares, permitindo a identificação de cromossomos e também uma maior compreensão sobre a evolução cariotípica, elucidando rearranjos e poliploidias (ÁVILA ROBLEDILLO et al., 2018; ČÍŽKOVÁ et al., 2013; KOO et al., 2011).

Em contraste à antiga suposição de sua falta de funcionalidade, vários DNAs repetidos em tandem aparentam estar ligados à complexas características cromossômicas em eucariotos, sendo componentes proeminentes da heterocramatina (PLOHL et al., 2008). Em termos de funcionalidade, os DNAs ribossomais (DNAr) 5S e 35S representam um dos grupos mais extensivamente estudados de DNA repetido em tandem. Isso se deve especialmente ao fato deles estarem sempre presentes no genoma eucarioto e por variarem extensivamente em número de sítiso e posição no cariótipo (WEISS-SCHNEEWEISS; SCHNEEWEISS, 2013). Além disto, os DNAr possuem alto valor em estudos evolutivos tanto por apresentarem relação com rearranjos cariotípicos (ROA; GUERRA, 2012) como por apresentarem regiões filogeneticamente informativas como os espaçadores transcritos interno e externo (ITS e ETS) (MARKOS; BALDWIN, 2001). Os telômeros (região terminal dos cromossomos eucariotos) representam outra região funcional cromossômica constituída por arranjos de sequências repetidas em tandem (WEISS-SCHNEEWEISS; SCHNEEWEISS, 2013). Apesar de ser conservada na maioria das angiospermas, estudos recentes têm demonstrado uma diversidade mais alta que o esperado nas sequências teloméricas de espécies das famílias Amaryllidaceae, Solanaceae e Lentibulariaceae (PESKA; GARCIA, 2020).

Outra região cromossômica funcional rica em elementos repetidos em tandem (bem como outros tipos de DNA repetitivo) é o centrômero (HOUBEN; SCHUBERT, 2003). Diferente das sequências teloméricas, o DNA repetitivo centromérico não é conservado entre grupos de plantas, levando a diversas questões sobre seu papel na evolução cariotípica (LEE et al., 2005). Apesar de ser um componente comum na maioria dos centrômeros dos eucariotos, alguns grupos de organismos não possuem DNA satélite específico do centrômero, o que implica que este não seja essencial para a formação do centrômero (HECKMANN; HOUBEN, 2013). Porém, estudos acerca da estrutura e formação de centrômeros têm mostrado que sequências de DNAs satélites podem promover propriedades físicas que facilitam a atividade de proteínas centroméricas, como a formação de nucleossomos (ESCUDEIRO et al., 2019; TSOUMANI et al., 2013). Neste contexto, a investigação dos diversos DNAs satélites associados às proteínas centroméricas têm sido de fundamental importância para uma melhor compreensão sobre a

evolução do centrômero e seu papel na diversificação genômica (MARQUES et al., 2015; PLOHL; MEŠTROVIĆ; MRAVINAC, 2014).

2.2.5 DNA repetitivo disperso

Apesar da existência de outros tipos de sequências repetitivas dispersas no genoma, os chamados elementos transponíveis são claramente os mais estudados (WEISS-SCHNEEWEISS; SCHNEEWEISS, 2013). Descobertos na década de 40 nos estudos pioneiros da vencedora do Nobel Barbara McClintock (MCCLINTOCK, 1948), os elementos transponíveis logo ganharam protagonismo em especial por sua marcante capacidade de se "mover" ao longo dos genomas hospedeiros. Tal característica faz dos elementos transponíveis os principais responsáveis pela variabilidade do tamanho do genoma, em especial nas angiospermas, onde chegam a constituir a grande maioria do genoma de certas espécies (ELLIOTT; GREGORY, 2015; GAUT; ROSS-IBARRA, 2008). Essas características levaram pesquisadores a acreditar que os elementos transponíveis eram "egoístas" e não possuíam nenhum valor adaptativo (SCHRADER; SCHMITZ, 2019). No entanto, isso não diminuiu o interesse no estudo destes, e dado a diversidade de diferentes elementos transponíveis, uma série de sistemas de classificação foram propostos ao longo dos anos (PIÉGU et al., 2015; WICKER et al., 2007).

Tradicionalmente, os elementos transponíveis foram divididos em duas grandes classes de acordo com seu método de replicação (FINNEGAN, 1989). Elementos de classe I, representados principalmente pelos retrotransposons, são transpostos por meio da ação de uma transcriptase reversa que cria cópias do elemento utilizando um intermediário de RNA. Elementos de classe II, dos quais os mais comuns são os transposons de DNA, se movimentam pelo genoma pela ação de transposases, responsáveis por cortar as sequências e reinseri-las em outras regiões. Partindo dessa classificação geral, vários outros sistemas têm sido propostos, levando em conta fatores como estrutura, similaridade e funcionalidade das sequências (ARKHIPOVA, 2017; KAPITONOV; JURKA, 2008; NEUMANN et al., 2019). Um tipo de retrotransposon particularmente interessante para a evolução das angiospermas são os LTR, assim chamados por possuírem um terminal repetido longo (LTR, do inglês *long terminal repeat*) flanqueando suas extremidades (HAVECKER; GAO; VOYTAS, 2004). Dentre os retrotransposons LTR, elementos pertencentes às linhagens *Ty3/Gypsy* e *Ty1*/*Copia* são encontrados principalmente na heterocromatina dos genomas das angiospermas, sendo grandes

responsáveis tanto pela composição de elementos cromossômicos funcionais como os centrômeros (MARQUES et al., 2015; ZHONG et al., 2002), como pela variação no tamanho do genoma (KELLY et al., 2015; MACAS et al., 2015; PIEGU et al., 2006).

Trabalhos recentes têm mostrado que o papel dos elementos transponíveis na evolução e diversificação biológica pode ser maior do que simplesmente promover "obesidade genômica" ou questões estruturais (SCHRADER; SCHMITZ, 2019). Estudos genômicos mais detalhados têm mostrado que a correlação entre tamanho do genoma e fatores ambientais como altitude e nicho ecológico é mediada pela amplificação/deamplificação de linhagens específicas de retrotransposons (BILINSKI et al., 2017; LYU et al., 2018; SCHLEY et al., 2021). Experimentalmente, já foi mostrado que diversos fatores como radiação UV, temperatura e tolerância à falta de água causam "explosões" de amplificação de elementos transponíveis (CAO; DENG; MCLAUGHLIN, 2014; KALENDAR et al., 2000; KIMURA et al., 2001; MATSUNAGA et al., 2015; RAMALLO et al., 2008). Somado a isso, tem sido especulado que a amplificação de elementos transponíveis em regiões gênicas ligadas à tolerância ambiental pode levar a um aumento de variabilidade genética em organismos sob pressão seletiva (GONZÁLEZ et al., 2010; SCHRADER et al., 2014). Em contrapartida, a perda de elementos transponíveis em espécies que migram para novos ambientes também tem sido reportada como uma possível estratégia de economia energética (LYU et al., 2018; PANDIT; WHITE; POCOCK, 2014). Não obstante, tem sido visto que a atividade de elementos móveis nos genomas hospedeiros pode ocasionar diversos processos adaptativos, como modificações na regulação gênica (HOF et al., 2016). Somada a estes fatores, a descoberta de que tanto a abundância (DODSWORTH et al., 2015) quanto à similaridade (VITALES; GARCIA; DODSWORTH, 2020) de elementos transponíveis possuem sinal filogenético continua a reforçar o possível papel evolutivo destes.

2.3 CROMOSSOMOS HOLOCÊNTRICOS

2.3.1 Estrutura e organização

Os centrômeros são estruturas cromossômicas caracterizadas pela presença de um complexo multi-proteico conhecido como cinetócoro, onde os microtúbulos se ligam durante a divisão celular, sendo essencialmente responsáveis pelo movimento cromossômico

(MELTERS et al., 2012). Na maioria dos eucariotos, o centrômero se localiza em uma região específica formando uma constrição primária no cromossomo, sendo estes definidos como monocêntricos. No entanto, alguns grupos não apresentam constrição primária, tendo seu centrômero disperso ao longo de quase todo o cromossomo, sendo conhecidos como holocêntricos (GUERRA et al., 2010). A diferença organizacional destes holocentrômeros implica em alguns obstáculos durante a divisão celular, mais especificamente na meiose (CUACOS; H. FRANKLIN; HECKMANN, 2015). Na meiose, dois eventos de segregação cromossômica (meiose I e II) procedem a replicação do DNA, onde, canonicamente, cromossomos homólogos são separados na meiose I e cromátides irmãs são separadas na meiose II (CABRAL et al., 2014). Durante a meiose I, cromossomos homólogos se unem formando bivalentes, facilitando a formação do quiasma de recombinação (GUERRA et al., 2010). Em monocêntricos, cinetócoros de cromátides irmãs se mantêm juntos, apontando para a mesma direção. Já em holocêntricos, a disposição do cinetócoro ao longo de quase toda a extensão das cromátides prejudicaria o direcionamento da segregação dos homólogos na meiose I, o que levou a evolução de diferentes modificações na meiose de organismos com esse tipo cromossômico (MELTERS et al., 2012). Notavelmente, algumas espécies holocêntricas apresentaram um tipo de meiose "invertida", onde as cromátides irmãs são separadas ainda em meiose I, separando-se de suas homólogas em meiose II (CABRAL et al., 2014; HECKMANN et al., 2014; LUKHTANOV et al., 2018; MARQUES et al., 2016).

Estruturalmente, os centrômeros, seja de cromossomos monocêntricos ou holocêntricos, apresentam dois componentes fundamentais: o DNA centromérico (altamente repetitivo) e proteínas centroméricas (histonas modificadas permanentemente ligadas ao DNA) (GUERRA et al., 2010). A variante histônica CenH3 aparece como a proteína responsável pelo estabelecimento do cinetócoro e pela manutenção do centrômero na maioria dos organismos estudados (CUACOS; H. FRANKLIN; HECKMANN, 2015). Apesar das divergências em nomenclatura (CenH3 nas plantas, CENP-A em mamíferos, CID em *Drosophila*) (EARNSHAW et al., 2013), a conservação de função da CenH3 ao longo da árvore da vida é marcante (GUERRA et al., 2010). Em contraste à essa conservação, as CenH3 apresentam rápida evolução de sequência, (HOUBEN; SCHUBERT, 2003) o que pode evidenciar que sua presença nos centrômeros tem um papel mais fundamental do que a sua sequência de DNA (GUERRA et al., 2010). Em cromossomos holocêntricos metafásicos, foi incialmente observado que a proteína centromérica se extende ao longo do lado externo das cromátides irmãs, tanto em animais (*Caenorhabditis elegans*) (MOORE; MORRISON; ROTH, 1999)

como em plantas (NAGAKI; MURATA, 2005). A observação destes holocentrômeros em uma maior resolução microscópica revela pequenos locos separados da proteína, dispersos ao longo da placa cinetocórica (STEINER; HENIKOFF, 2014). Em algumas espécies de plantas com cromossomos holocêntricos grandes, como *Rhynchospora pubera* (MARQUES et al., 2015) e *Luzula elegans* (HECKMANN et al., 2011), foi observado que a CenH3 forma um sulco longitudinal nas cromátides de cromossomos metafásicos. No entanto, ainda existe dúvidas se esse sulco centromérico é uma adaptação relacionada a cromossomos grandes ou uma novidade evolutiva de certos gêneros (CUACOS; H. FRANKLIN; HECKMANN, 2015).

2.3.2 DNA repetitivo centromérico

Diferente do alto nível de conservação de outras regiões cromossômicas funcionais como os telômeros, o DNA centromérico apresenta uma marcante diversidade de sequências entre os organismos estudados (GUERRA et al., 2010). Em aspectos funcionais, o "centrômero-ponto" das leveduras foi até hoje o único a apresentar uma sequência de DNA específica essencial para a localização e formação da proteína CenH3 (CLARKE; CARBON, 1985). Em todos os outros eucariotos investigados, o DNA centromérico é composto por arranjos de milhares de bases de DNA repetitivo, especialmente DNAs satélites e elementos transponíveis (PLOHL; MEŠTROVIĆ; MRAVINAC, 2014). Essa falta de especificidade levanta a hipótese de que sequências de DNA centromérico não são suficientes nem necessárias para a formação do centrômero funcional (HECKMANN; HOUBEN, 2013).

No contexto da investigação das diversas sequências de DNA repetitivo centromérico, técnicas de imunoprecitação de cromatina (ChIP) com a CenH3 têm sido essenciais para o isolamento e sequenciamento de sequências presentes nas regiões centroméricas (PLOHL; MEŠTROVIĆ; MRAVINAC, 2014). Em espécies monocêntricas, várias famílias de DNA satélite linhagem-específicas foram descobertas, tais como os satélites alfa em humanos (WILLARD; WAYE, 1987), satélites AATAT e AAGAG em *Drosophila* (SUN; WAHLSTROM; KARPEN, 1997), pAL1 em *Arabidopsis thaliana* (NAGAKI et al., 2003), *CentC* em *Zea mays* e em várias outras gramíneas (PLOHL; MEŠTROVIĆ; MRAVINAC, 2014) e 13 diferentes famílias em *Pisum sativum* (NEUMANN et al., 2011). Quanto aos elementos transponíveis, destacam-se retrotransposons LTR *Ty3-Gipsy* do tipo Cromovírus, assim chamados por possuírem um cromodomínio (NEUMANN et al., 2011). Entre estes, estão os CRM1 e CRM2 do milho (ZHONG et al., 2002), CRR em *Oryza sativa* (CHENG et al.,

2002), CRS em *Saccharum officinarum* (NAGAKI; MURATA, 2005), dentre outros. De acordo com seu agrupamento filogenético, todos estes elementos Cromovírus são usualmente referidos como elementos do clado CRM (NEUMANN et al., 2011). Já em holocêntricos, acreditou-se por muito tempo que não existiam sequências de DNA satélite ou retrotransposons específicos do centrômero (GUERRA et al., 2010). Experimentos de ChIP com a CenH3 tanto no nematódeo *C. elegans* (STEINER; HENIKOFF, 2014) como na planta *L. elegans* (HECKMANN et al., 2011) não identificaram nenhuma sequência de DNA centrômero-específica. No entanto, Marques et al. (2015) identificaram em *Rhynchospora pubera* a família de DNA satélite *Tyba* e um retrotransposon da linhagem dos cromovírus (denomidado CRRh), sendo os primeiros elementos repetitivos centrômero-específicos descobertos em uma espécie holocêntrica.

2.3.3 Cromossomos holocêntricos em *Rhynchospora* Vahl. (Cyperaceae)

Acredita-se que cromossomos holocêntricos apareceram independentemente quatro vezes ao longo da evolução das angiospermas, sendo uma destas supostamente em um ancestral comum das famílias Cyperaceae e Juncaceae, da ordem Poales (MELTERS et al., 2012). Junto à família Thurniaceae, estas formam um clado informal conhecido como Cyperid (CHASE et al., 2006; LINDER; RUDALL, 2005). Apesar de acreditar-se que todas es espécies de Cyperid possuíam cromossomos holocêntricos, estudos recentes confirmaram a presença de espécies monocêntricas em Thurniaceae (BAEZ et al., 2020) e Juncaceae (GUERRA; RIBEIRO; FELIX, 2019). Cyperaceae é de longe a família mais diversa do grupo, com aproximadamente 5500 espécies divididas em cerca de 98 gêneros, configurando a décima família com maior número de espécies em angiospermas (GOVAERTS; SIMPSON, 2007). Estima-se que a família tenha se originado na América do Sul no período Cretáceo, apesar de seus clados mais diversificados estarem presentes nas regiões temperadas (SPALINK et al., 2016). Dentre todos os grupos com presença confirmada de cromossomos holocêntricos, Cyperaceae é de longe o mais diverso, com um bom número de espécies em vários gêneros sendo confirmadas como holocêntricas (ESCUDERO; WEBER; HIPP, 2013; ROALSON, 2008; ROALSON; MCCUBBIN; WHITKUS, 2007).

Dentro desta família, espécies com cromossomos maiores como algumas encontradas no gênero *Rhynchospora* permitem um estudo mais minucioso da estrutura dos holocentrômeros em Cyperaceae (CABRAL et al., 2014; MARQUES et al., 2015, 2016; RIBEIRO et al., 2017).

O gênero *Rhynchospora* possui aproximadamente 400 espécies com uma distribuição cosmopolita, representando um dos gêneros mais diversos de Cyperaceae (THOMAS; ARAÚJO; ALVES, 2009). O gênero apresenta sua origem estimada em aprox. 45 milhões de anos atrás na América do Sul, com posterior dispersão e diversificação no hemisfério norte, especialmente na América do Norte (BUDDENHAGEN, 2016). Filogeneticamente, o grupo apresentou por muito tempo classificações infragenéricas confusas, com diferentes sistemas de classificação tribal propostos (GALE, 1944; KÜKENTHAL, 1939). Recentemente, filogenias robustas do grupo têm sido propostas, resolvendo vários problemas de classificação, como por exemplo a sinonimização do antigo gênero *Pleurostachys*, agora parte de *Rhynchospora* (BUDDENHAGEN, 2016; THOMAS, 2020). Cariotipicamente, poliploidias e disploidias foram observadas como eventos importantes para a evolução cromossômica do grupo, com uma ampla variabilidade de tamanhos do genoma também observada (BURCHARDT et al., 2020; RIBEIRO et al., 2018)

Estudos citogenômicos recentes em *Rhynchospora* têm apresentado várias novidades quanto ao comportamento e estrutura de cromossomos holocêntricos. Tanto *R. pubera* como *R. tenuis* também serviram como modelo para confirmar o modelo de meiose invertida de *Rhynchospora*, onde as cromátides irmãs são separadas em meiose 1, ligadas por eucromatina às cromátides de cromossomos homólogos em prófase 2 e subsequentemente separadas na anáfase 2 (CABRAL et al., 2014). Uma análise mais detalhada em *R. pubera* mostrou uma diferença na organização das unidades centroméricas em mitose e meiose, configurando um caso inédito de reorganização do centrômero durante a divisão celular (MARQUES et al., 2016). Em *R. pubera* foi observado que a proteína centromérica CenH3 estava disposta em um sulco longitudinal ao longo da cromátide (MARQUES et al., 2015), assim como tinha sido observado em *Luzula elegans* (HECKMANN et al., 2011). Também em *R. pubera* foram descobertos o DNA satélite *Tyba* e o retrotransposon CRRh, ambos co-localizados com a CenH3, configurando os primeiros DNAs repetitivos específicos do centrômero de uma espécie holocêntrica (MARQUES et al., 2015). Além de serem os primeiros elementos repetitivos centrômero-específicos de holocêntricos, a disposição de *Tyba* e CRRh intercalados à eucromatina ao longo dos sulcos centroméricos configuou uma forma de organização centromérica até então inédita (MARQUES et al., 2015). É interessante notar que *Tyba*, além de ser a família de DNA satélite mais abundante em *R. pubera*, foi também identificada posteriormente como mais abundante em duas outras espécies do gênero (*R. ciliata* e *R. tenuis*), novamente apresentando localização centrômero-específica (RIBEIRO et al., 2017).

Notavelmente, neste mesmo trabalho, *Tyba* não foi encontrado em *R. globosa*, mostrando que a presença de um único satélite centrômero-específico não está conservada no gênero. Estas características fazem de *Rhynchospora* um modelo potencialmente interessante para o estudo da evolução de DNAs satélites e outros elementos repetitivos em um contexto macroevolutivo.

2.4 MÉTODOS COMPARATIVOS FILOGENÉTICOS

2.4.1 Introdução

Estudar a variabilidade de caracteres e traços funcionais de grupos de organismos é um dos pilares das ciências biológicas e representa uma etapa fundamental na compreensão da diversidade biológica (UYEDA; ZENIL-FERGUSON; PENNELL, 2018). No entanto, o significado evolutivo da variação de caracteres em um grupo de organismos só pode ser investigado sob a luz de uma árvore filogenética, que pode ser definida como a representação do histórico de separação das espécies ao longo do tempo (O'MEARA et al., 2016). Uma árvore filogenética possui nós, que representam ancestrais hipotéticos, e ramos, que podem variar de tamanho em unidades de tempo, número de gerações ou mudanças ao longo do tempo (O'MEARA et al., 2016). Associando valores de caracteres aos terminais de ramos de uma árvore filogenética, é possível levar em conta a "não-independência" entre as histórias evolutivas do caractere e do grupo estudado (HARMON, 2019). Neste contexto, os métodos comparativos filogenéticos (MCFs) apresentam uma elegante solução interdisciplinar, juntando conceitos da biologia, paleontologia, ecologia e matemática na busca da compreensão sobre a evolução e diversificação das espécies (PENNELL; HARMON, 2013).

A necessidade dos MCFs para o estudo da evolução de caracteres foi primeiramente apontada por Felsenstein (1985). Neste trabalho, Felsenstein se inspirou em uma propriedade física que explica o movimento aleatório de partículas ao longo de determinado tempo, conhecida como movimento Browniano. Ao assumir que um caractere evolui aleatoriamente ao longo do tempo (e que, quanto maior esse tempo, maior a mudança) é possível estimar o valor do caractere em determinado período (HARMON, 2019). Como o tamanho dos ramos de uma árvore filogenética reflete o tempo de divergência entre espécies, é possível usar essa informação para reconstruir ou traçar correlações entre caracteres ao longo da evolução de um grupo utilizando os princípios do movimento Browniano (FELSENSTEIN, 1985).

A princípio, os MCFs surgiram como ferramentas estatísticas que tinham apenas o objetivo de lidar com o "obstáculo" da ancestralidade compartilhada representada em árvores ao se estudar um determinado caractere (HARVEY; PAGEL, 1991). No entanto, avanços computacionais têm permitido a criação de meios cada vez mais precisos e complexos para a criação e datação de árvores filogenéticas (DRUMMOND; RAMBAUT, 2007; RONQUIST; HUELSENBECK, 2003; STAMATAKIS, 2006). Como consequência destes avanços, um alto número de novos MCFs têm sido criados para o estudo dos mais diversos caracteres e fenômenos (O'MEARA, 2012; PENNELL; HARMON, 2013). Particularmente, tem sido observado um aumento no uso de MCFs para o estudo de caracteres citogenéticos e genômicos, em busca de uma maior compreensão acerca da evolução destes caracteres, bem como seu impacto na distribuição e diversificação de espécies (COSTA et al., 2017, 2020; RIBEIRO et al., 2018; SADER et al., 2019; SOUZA et al., 2019).

Ao estudar a evolução de um ou mais caracteres ao longo de uma filogenia, os valores estimados para os ancestrais teóricos (representados pelos nós da árvore) são essenciais para a investigação do impacto destes caracteres no histórico de separação de linhagens (GOOLSBY, 2017). A base da reconstrução de caracteres contínuos ao longo de uma filogenia está atrelada ao conceito de contrastes independentes filogenéticos (PICs, do inglês *Phylogenetic independente contrasts*) introduzido por Felsenstein (1985). Os PICs são calculados de acordo com a mudança de um caractere em linhagens irmãs em função do tempo de divergência entre estas, possibilitando a estimativa de um índice de evolução deste caractere (HARMON, 2019). A simplicidade do princípio dos PICs para a reconstrução de caracteres permite sua aplicação em diversas áreas. Notavelmente, a reconstrução de caracteres contínuos citogenéticos como o tamanho do genoma (COSTA et al., 2017; SOUZA et al., 2019) e genômicos, como abundância de DNA repetitivo (KELLY et al., 2015; MACAS et al., 2015) têm sido fundamental para uma maior compreensão de como estes caracteres tão variáveis impactam a evolução de diferentes grupos.

2.4.2 Taxa de diversificação

Com a ascensão do "pensamento em árvore" e dos MCFs, pesquisadores começaram a perceber que árvores filogenéticas contém informações sobre padrões históricos de diversificação de espécies (PENNELL; HARMON, 2013). Os métodos mais simples para se estimar taxa de diversificação em uma árvore levam em conta apenas a diferença em número

de espécies entre clados irmãos (SLOWINSKI; GUYER, 1993). Normalmente estes métodos comparam as taxas de diversificação obtidas com um modelo nulo de nascimento-morte onde todas as linhagens têm a mesma probabilidade de especiação e extinção (PENNELL; HARMON, 2013). Esta abordagem, no entanto, possui algumas limitações, visto que taxas de diversificação são bem mais dinâmicas do que simples modelos nascimento-morte e bastante dependentes do tamanho dos ramos, ou seja, do tempo de diversificação das linhagens (ALROY, 2010). No entanto, avanços nas ferramentas bioinformáticas têm proporcionado a inclusão de parâmetros dependentes de tempo em análises de diversificação, bem como a rápida comparação de modelos por verossimilhança (CHAN; MOORE, 2005) ou inferência Bayesiana (RABOSKY, 2014). Combinado a análises de reconstrução de caracteres ancestrais, é possível estimar possíveis correlações entre transição de caracteres e diversidade de espécies em diferentes grupos. Notavelmente, esta relação foi observada em angiospermas para uma série de caracteres como sistema reprodutivo (SABATH et al., 2016) e simetria floral (SARGENT, 2004). Recentemente, mudanças a níveis cromossômicos também têm sido colocadas como possíveis motores de diversificação, tais como poliploidia (SADER et al., 2019; TANK et al., 2015) e surgimento de cromossomos holocêntricos (MÁRQUEZ-CORRO; ESCUDERO; LUCEÑO, 2018).

2.4.3 Análises de correlação

Testes de correlação entre diferentes caracteres são comuns para testes de hipóteses em estudos comparativos (HARMON, 2019). Ao usar MCFs para levar em conta a filogenia, é possível entender não apenas como o valor de um caractere afeta outro, mas também como a direção e magnitude de mudanças ao longo da evolução destes caracteres podem estar relacionadas (PENNELL; HARMON, 2013). O método de PICs de Felsenstein (1985) foi primeiramente utilizado para a correlação de caracteres ao longo da filogenia, sendo até hoje bastante popular para este fim. Outro método de correlação análogo aos PICs é o método de mínimos quadrados generalizados filogenéticos (PGLS, do inglês *Phylogenetic generalized least squares*) (GRAFEN, 1989). Apesar da semelhança com PICs, o método de PGLS é mais flexível permite o teste de modelos de evolução diferentes do movimento Browniano (PENNELL; HARMON, 2013). Outra possibilidade para ajustar análises de correlação a um contexto filogenético, é o uso de autovetores filogenéticos (DINIZ-FILHO; DE SANT'ANA; BINI, 1998). Esse método baseia-se na obtenção de autovetores a partir de matrizes de distância filogenética, atribuindo valores específicos aos terminais que retratam proximidade

filogenética, podendo ser usados como variáveis em análises de correlação (DINIZ-FILHO; DE SANT'ANA; BINI, 1998; GUÉNARD; LEGENDRE; PERES-NETO, 2013). Estes métodos de correlação têm sido importantes ferramentas na citogenética, permitindo por exemplo correlacionar tamanho do genoma, número cromossômico e fatores ambientais em grupos marcados por poliploidia e/ou variação heterocromática (COSTA et al., 2017, 2020; SADER et al., 2019; VAN-LUME et al., 2017).

É importante ressaltar que estes métodos possuem a limitação de permitirem apenas um valor de caractere por terminal da árvore (geralmente a média ou mediana), dificultando a investigação de caracteres que podem ser informativos em diferentes indivíduos, tais como variáveis ambientais (ADAMS; COLLYER, 2018). Além disso, análises simples de PICs e PGLS são eficientes para testes de correlação entre dois caracteres, mas não possuem aplicabilidade para análises multivariadas (PENNELL; HARMON, 2013). Felizmente, pesquisadores tem desenvolvido uma grande diversidade de MCFs para testes de correlação multivariados (revisado por Adams & Collyer, 2018). Com a atual facilidade para obtenção de dados de distribuição geográfica e variáveis ambientais, um dos principais usos destes métodos tem sido a correlação entre diferentes caracteres e a plasticidade ecológica. Como exemplo, trabalhos recentes têm usado estes métodos para avaliar a correlação entre caracteres como o tamanho do genoma e variáveis relacionadas ao solo, temperatura e precipitação (GUIGNARD et al., 2016; SCHLEY et al., 2021; SOUZA et al., 2019).

2.4.4 Filogenômica Comparativa

Com a era genômica, a busca pela integração de ferramentas capazes de processar e analisar a alta quantidade de dados de sequência gerados têm sido uma constante dentro das ciências biológicas (DODSWORTH et al., 2015). Relações filogenéticas, antes inferidas com o auxílio de algumas poucas regiões do DNA, passaram a ser construídas mediante análise de complementos genômicos completos, tais como os de mitocôndrias (VAN DE PAER et al., 2016), plastídios (BARRETT et al., 2013) e até mesmo de todo o conjunto de genes codificantes (YODER et al., 2013). Além de conferir maior robustez às relações filogenéticas, a grande quantidade de dados gerados por estes métodos tem permitido a análise de outros caracteres genômicos em um contexto evolutivo por meio de análises derivadas de MCFs (DODSWORTH et al., 2015; LANG et al., 2010; OAKLEY et al., 2005; VITALES; GARCIA; DODSWORTH, 2020).

Em angiospermas, a fração repetitiva do DNA constitui um dos caracteres genômicos mais estudados, principalmente por sua grande diversidade e abundância na maioria das espécies, além de seu impacto na variabilidade de tamanho do genoma observada no grupo (MACAS et al., 2015; PLOHL et al., 2008; WEISS-SCHNEEWEISS et al., 2015). A possibilidade de sequenciar com alta resolução a fração repetitiva por meio de sequenciamentos de baixa cobertura (menos de 1% do genoma) como *genome skimming* deslanchou os estudos sobre DNA repetitivo (STRAUB et al., 2012). Somado a isso, a contínua criação de bases de dados e ferramentas para a identificação e caracterização de elementos repetitivos têm sido essenciais para o estudo destes em um contexto comparativo. Notavelmente, a plataforma online do RepeatExplorer permite o agrupamento de elementos repetitivos por similaridade, estimando a abundância genômica de diferentes classes de DNA repetitivo (NEUMANN et al., 2019; NOVAK et al., 2013). Com o uso destas ferramentas, Dodsworth et al. (2015) compararam a abundância de diferentes classes de elementos repetitivos entre diferentes espécies e descobriram que estas abundâncias, tratadas como caracteres contínuos, possuíam sinal filogenético. Dessa forma, neste estudo e em outros subsequentes, foi possível usar a abundância de elementos repetitivos para gerar hipóteses filogenéticas entre espécies (DODSWORTH et al., 2015, 2016, 2017). Recentemente, este mesmo grupo de pesquisa apresentou uma outra forma de utilizar elementos repetitivos para traçar relações filogenéticas, desta vez comparando matrizes de similaridade das diferentes classes de elementos repetitivos (VITALES; GARCIA; DODSWORTH, 2020). Tendo em vista estes novos métodos, a investigação dos elementos repetitivos em um contexto filogenético é essencial para a busca pelo seu significado evolutivo e sua influência na diversidade das angiospermas.

3 RESULTADOS

3.1 WHAT DETERMINES RATES OF SATDNA EVOLUTION? INVESTIGATING THE DIVERSIFICATION OF HOLOCENTROMERIC SATDNA *TYBA* IN *RHYNCHOSPORA* (CYPERACEAE)

*Artigo submetido à revista Annals of Botany (https://academic.oup.com/aob)

Original Article

What determines rates of satDNA evolution? Investigating the diversification of holocentromeric satDNA *Tyba* in *Rhynchospora* (Cyperaceae)

Lucas Costa[1]*, André Marques[2], Chris Buddenhagen[3], Andrea Pedrosa-Harand[1], Gustavo Souza[1]

[1] *Laboratory of Plant Cytogenetics and Evolution, Department of Botany, Federal University of Pernambuco, Recife-PE, Brazil*

[2] *Department of Chromosome Biology, Max Planck Institute for Plant Breeding Research, Cologne, Germany*

[3] *AgResearch, Plant Functional Biology, Ruakura, New Zealand*

Running Title: Investigating the diversification of the satDNA *Tyba* in *Rhynchospora*

***** lucas.costa.18@hotmail.com

ABSTRACT

- *Background and Aims*: *Satellite* DNAs (satDNAs) are repetitive sequences composed by tandemly arranged, often highly homogenized units called monomers. Although satDNAs are usually fast evolving, some satDNA families may be conserved across species separated by several millions of years, related to their functional roles in the genomes. *Tyba* was the first centromere-specific satDNA described for a holocentric organism, until now being characterized for only few species of the *Rhynchospora* Vahl. (Cyperaceae). Here, we investigated the evolution of satDNAs across the *Rhynchospora* genus.

- *Methods*: We characterized structure and sequence evolution of satDNAs across a robust hyb-seq-based dated phylogeny of 70 species. We mined the repetitive fraction for *Tyba*-like satellites to compare its features to other satDNAs and to construct a *Tyba*-based phylogeny for the genus.

- *Key Results*: Our results show that *Tyba* is present in most of the genus, spanning four out of five major clades and maintaining intrafamily pairwise identity of 70.9% over 31 My. In comparison, other canonical satellite families present higher intrafamily pairwise identity, but are phylogenetically restricted. Furthermore, *Tyba* sequences could be divided in 12 variants grouped into three different clade-specific subfamilies, showing evidence of traditional models of satDNA evolution, such as concerted evolution and library hypothesis. Besides, a *Tyba*-based phylogeny showed high congruence with the hyb-seq topology. We suggest that *Tyba* has a specific interaction with nucleosomes, given its high curvature peaks over conserved regions and overall high bendability values compared to other non-centromeric satellites.

- *Conclusions*: Overall, our results show a remarkable sequence conservation and phylogenetic significance for *Tyba* across the genus *Rhynchospora*, which suggests that functional roles may lead to long-time stability and conservation for satDNAs in the genome.

Keywords: Holocentromere, Repetitive DNA, sateliite DNA, Phylogenetic signal, *Rhynchospora* Vahl.

INTRODUCTION

Repetitive DNA has become a key element in genomic studies since it was discovered that these sequences compose a large fraction of most eukaryotic genomes (Gemmell, 2021). The satellite DNAs (satDNAs) are one of the most well-studied types of repetitive DNA, being composed by tandemly arranged, usually highly homogenized units called monomers (Garrido Ramos, 2017). Monomer consensus sequences are used to characterize different families of satDNA, which can be numerous even considering a single species (Novák *et al*., 2017; Oliveira *et al*., 2021). One of the key features of these sequences is the remarkably fast rates of sequence evolution, in terms of both abundance and nucleotide sequence, resulting in accumulation of changes in short evolutionary times (Macas *et al.* 2002; Lower *et al.* 2017). This rapid diversification present major challenges to study satDNA evolution (Plohl *et al*., 2012).

Although many theories have been proposed to explain satDNA evolution (Lower *et al*., 2017), the most prominent models are the concerted evolution and library model, which are often complementary to each other (Plohl *et al*., 2012; Garrido-Ramos, 2015; Camacho *et al*. 2021). The concerted evolution model follows the idea that the monomers of a satDNA array accumulate mutations in an independent manner and that these mutations are homogenized and fixed along the array following a molecular drive process (Dover, 2002; Plohl *et al*., 2012; Garrido-Ramos, 2015). The library model postulates that related species share a library of satDNA families of varying abundances caused by random expansion or contraction of arrays of these different satDNA (Salser *et al*., 1976; Fry and Salser, 1977).

While these two models try to explain the fast changes of these repeats in sequence and abundance, there are cases of long-time satDNA sequence conservation, usually observed in higher taxonomic levels such as families (Mravinac *et al*., 2002), orders (Robles *et al*., 2004) and even phylum (Petraccioli *et al*., 2015). The maintenance of "relic" satDNA families may be indicative of functional roles for these sequences in eukaryotic genomes (Plohl *et al*., 2012).

This functionality may be related to expression of non-coding RNA sequences, associated with nucleotypic and/or structural effects (Mravinac *et al*., 2005). Sequence-based nucleosome-prediction models using conserved satDNA monomers suggest a role to facilitate nucleosome formation (Tsoumani *et al*., 2013; Escudero *et al*., 2019). These models analyse sequence properties such as dinucleotide periodicity and curvature patterns to predict 'bendability' values of a DNA sequence, which can be an indicator of interaction with histones (Liu *et al*., 2011; Zhang *et al*., 2013).

SatDNAs are usually an integral part of the heterochromatin, being frequently found in functional regions such as the telomere and centromere (Achrem *et al*., 2020). In regard to the centromere, most eukaryotes present monocentric chromosomes, in which the centromere is restricted to a single region, usually composed by long arrays of repetitive sequences, especially satDNAs (Plohl *et al*., 2014). These centromere-specific satDNAs present varying degrees of sequence conservation (Melters *et al*. 2013). For example, within the Poaceae family, the CentO centromeric satDNA found in *Oryza* was later shown to have high similarity with CentC, present in the centromeres of maize (Cheng *et al*., 2002; Zhong *et al*., 2002). In contrast, centromeric satDNAs of some *Solanum* species were shown to be chromosome specific (Gong *et al*., 2012; Zhang *et al*., 2014). In holocentric species, which present a dispersed centromere along each chromatid (Bureš *et al*., 2013), it was believed for a long time that centromere-specific repetitive sequences were not present (Marques and Pedrosa-Harand, 2016). For example, well-studied holocentric organisms such as *Luzula elegans* and *Caenorhabditi elegans* did not present repeats associated with the centromeric protein, even after detailed genomic characterizations (Subirana and Messeguer 2013, Heckmann *et al.* 2011). This changed with the discovery of the 172-bp satDNA *Tyba* in the sedge species *Rhynchospora pubera* (Cyperaceae). *Tyba* showed a line-like distribution along a grove positioned at the outer part of chromatids in *R. pubera* chromosomes co-localizing with the CENH3 protein, being the

first centromere-specific satDNA described for a holocentric species (Marques *et al*., 2015). *Tyba* sequences were later discovered in four other *Rhynchospora* species (*R. breviuscula, R. cephalotes, R. ciliata, R. exaltata* and *R. tenuis*), while being absent in *R. globosa* (Rocha *et al.* 2016, Ribeiro *et al.* 2017, Costa *et al*. 2021). This already suggests an old origin for this satellite DNA, given the estimated distance (approx. 30 My, Buddenhagen, 2016) between the clade containing *R. cephalotes* + *R. exaltata* and the clade containing the other species that presented *Tyba*. Whole genome sequencing of *Rhynchospora* species revealed that *Tyba*-based holocentromeres also impact genomic architecture, epigenome organization, and karyotype evolution (Hofstatter *et al*. 2022). This suggests that the presence of *Tyba* at centromeres may have an adaptive role and may influence species diversification.

Rhynchospora is a cosmopolitan genus comprised of approx. 400 species with a north American centre of diversity, with preliminary divergence time estimates placing the origin of the genus between 38 and 49 Mya (Buddenhagen, 2016; Thomas, 2020; Silva Filho *et al*., 2021). The presence of *Tyba* across evolutionary distant species, coupled with its holocentromeric localization, makes this sequence an interesting case to study the evolution and dynamics of centromeric DNA in non-monocentric organisms in a macroevolutionary context. Here, we perform a genus-wide investigation about the tempo and mode of *Tyba* evolution in *Rhynchospora*, comparing it with other satDNAs found in the genus. We mined repetitive DNA information from filtered *off-target* NGS reads (Costa *et al.* 2021) from 70 species of *Rhynchospora* representing the major clades of the genus and used a robust phylogenetic framework to serve as background for studying *Tyba* evolution. We aimed to answer the following questions: 1) How widely distributed is *Tyba* in *Rhynchospora* when compared to other satDNAs? 2) How conserved is *Tyba* across the whole genus? 3) Is the evolutionary persistence of *Tyba* related to its centromeric distribution?

METHODS

Sequence data acquisition and filtering

All target-capture sequencing data analysed here were obtained from Buddenhagen (2016). Because we used off-target reads from target capture sequencing, we opted to exclude from our analysis the data of any species that showed a percentage of annotated repeats smaller than the one obtained for *R. cephalotes* by Costa *et al.* (2021), which was the species with less classified repeats (~5%) that still presented good correlation values with genome skimming data. In total, 77 accessions representing 70 *Rhynchospora* Vahl species (~20% of the genus) were selected for our satellite mining analysis. From the same dataset, we collected data from six *Carex* L. species, two *Chorizandra* R. Br. species, *Exocarya sclerioides* Benth., *Hypolytrum nemorum* (Vahl) Spreng. and *Scirpodendron ghaeri* (Gaertn.) Merr. to serve as outgroup **[Supplementary Table 1]**. All sequences used were deposited on GenBank under project number PRJNA672127.

Phylogenetic analyses and molecular dating

To anchor our findings in a phylogenetic backbone, we used the robust RaxML topology constructed by Buddenhagen (2016) with 256 target loci obtained by hybrid target-capture sequencing. Although the original sampling contained 115 *Rhynchospora* accessions, the reads of some of these were not sufficient for the RepeatExplorer analysis, yielding poor annotations. Therefore, we pruned the original tree leaving only 77 *Rhynchospora* accessions and the 11 outgroup species. This was done with the *drop.tip* function implemented in the package *phytools* (Revell, 2012) in Rstudio (R Core Team, 2019). This pruned tree was then submitted to a molecular clock analysis. Divergence times were estimated on BEAST v.1.8.3 (Drummond and Rambaut, 2007) through CIPRES Science Gateway using the pruned tree as fixed topology. For calibration, we used the same points defined by Buddenhagen (2016) following a normal

distribution with a 10% standard deviation. An uncorrelated relaxed lognormal clock (Drummond and Rambaut, 2007) and Birth-Death speciation model (Gernhard, 2008) were applied. Two independent runs of 100,000,000 generations were performed, sampling every 10,000 generations. After removing 25% of samples as burn-in, the independent runs were combined and a maximum clade credibility (MCC) tree was constructed using TreeAnnotator v.1.8.2 (Rambaut and Drummond, 2013). In order to verify the effective sampling of all parameters and assess convergence of independent chains, we examined their posterior distributions in TRACER. The MCMC sampling was considered sufficient at effective sampling sizes (ESS) equal to or higher than 200.

Satellite DNA mining

In order to use the target-capture sequencing data for satDNA mining, we first had to filter out all the reads containing the enriched targeted regions. For this, we follow the protocol presented by Costa *et al*. (2021) to acquire the unenriched off-target portion of the genomic libraries. The off-target datasets of each *Rhynchospora* species were uploaded to the RepeatExplorer pipeline (Novak *et al*., 2013) hosted at the web-based platform Galaxy (https://repeatexplorer-elixir.cerit-sc.cz/). RepeatExplorer uses a graph-based clustering algorithm to group sequences based on similarity, facilitating the identification of high copy sequences of a genome. These clusters of sequences are then identified by cross-checking against repetitive element databases (Novak *et al*., 2013). Concurrently, the TAREAN (Tandem Repeat Analyzer) tool checks the clusters for predictions of tandem arrangement, building consensus sequences for these clusters (Novák *et al*., 2017).

Our dataset was submitted to three different run strategies in RepeatExplorer: i) individual species clustering (ISC) analysis for each of the 77 *Rhynchospora* accessions in order

to characterize the consensus of the satDNAs of each species; ii) a comparative clustering (CC) analysis with reads from all the *Rhynchospora* accessions, in order to identify shared satellite DNAs and iii) A *Tyba*-like clustering (TLC) analysis, a comparative analysis using only reads that were mapped to a database of previously published *Tyba* using the *Geneious read mapper* implemented in Geneious v 7.1.9 (low sensitivity preset, Kearse *et al*., 2012). Since genome sizes of most analysed species were unknown, we inputted all reads left after the filtering of target-regions (off-target reads) for the ISC analysis, while the same number of reads (170,000) for each species was used to build the combined date set for the CC analysis. This amount was decided based on the species that had the smallest number of reads analysed in the individual analysis (*R. glaziovii*, 173,544; **Supplementary Table 1**). For the TLC analysis, we used all the 157,078 reads that were mapped to the *Tyba* database. The run parameters of the individual and comparative RepeatExplorer analysis were the same as Costa *et al*. (2021).

The CC and TLC analysis were used to divide *Tyba* into subfamilies and variants respectively **[Supplementary Figure 1]**. The consensus sequences of clusters annotated as *Tyba* in the CC analysis were aligned using the Geneious alignment tool with default settings (Kearse *et al*., 2012). The alignment was used to produce an "Approximately-Maximum-Likelihood" (AML) phylogenetic tree using FastTree, as implemented in Geneious v.7.1.9 (Kearse *et al*., 2012), forming monophyletic clades that were considered different subfamilies of the satDNA **[Supplementary Figure 1]**. The similarity between the shared satDNA found in the CC analysis (including non-*Tyba* satDNA families) was also assessed by a dotplot constructed on DOTTER (Sonnhammer and Durbin, 1995). The consensus sequences found in the TLC analysis were also used to construct an AML tree, and each of the clusters could be mapped to one of the subfamilies, being considered variants **[Supplementary Figure 1]**. The consensus sequences found in the ISC analysis were mapped to the *Tyba* variants and non-*Tyba* shared satDNAs to estimate sequence diversity of each shared satDNA. To assess the

relationship between sequence divergence and age of the shared satDNAs, the pairwise identity (%) of sequences mapped to each shared satDNA was estimated on Geneious and plotted against the time of divergence between species that contained each of the shared satDNA.

Satellite DNA-based phylogenetic analysis

We also performed a phylogenetic analysis based on the satDNA family *Tyba* and a separate analysis based on the other satDNAs found in the genus. For this, we separated all reads annotated as *Tyba* from the reads identified as other satDNAs ("non-*Tyba*" satDNAs) in the ICS analysis and employed an Alignment and Assembly Free (AAF) methodology (Fan *et al*., 2015). AAF constructs phylogenies directly from unassembled genome sequence data, bypassing both genome assembly and alignment. Thus, it calculates the statistical properties of the pairwise distances between genomes, allowing it to optimize parameter selection and to perform bootstrapping.

We also used the individual *Tyba* consensus from the ICS analysis to estimate the phylogenetic signal (Pagel's λ, Pagel 1999) of monomer length and GC content of *Tyba* sequences. For this, we use the *phytools* package (Revell, 2012) implemented in Rstudio (R Core Team, 2019). For species that had more than one *Tyba* variant, we used the consensus sequence of the most abundant *Tyba* cluster **[Supplementary Table 2]**. This analysis presents a measure of similarity between phylogenetic close species regarding the studied characteristics. In this case, a value of λ closer to 1 would mean that closely related species would have more similar *Tyba* variants, whereas λ closer to 0 would mean that closely related species had less similar *Tyba* variants than expected (Pagel 1999).

Sequence conservation

The consensus sequence of the 532 satDNAs found in individual clustering analysis were mapped to the consensus sequences of the shared satDNAs revealed in the comparative clustering analysis using the BowTie2 mapper (high sensitivity preset, End-to-End) (Langmead and Salzberg, 2012). We used a 60% pairwise identity threshold to determine whether a group of satellite DNAs belonged to the same family. The resulting alignment was used to calculate pairwise identity between all sequences mapped to each satDNA. The age of shared satellites was estimated based on the age of the most common recent ancestor (MRCA) between the species that possessed the satDNAs and a plot against pairwise identity was constructed using RStudio.

Curvature/bendability analysis of shared satDNAs

Consensus sequences of the *Tyba* variants (mapped to their respective subfamily consensus sequences) and of the other shared satDNAs (including the non-centromeric *Rg*Sat) were used to estimate bendability and curvature plots based on DNA sequence using the bend.it server (http://pongor.itk.ppke.hu/dna/bend_it.html). It uses the DNase I based bendability parameters of Brukner *et al.* (1995) and the consensus bendability scale (Gabrielan and Zohary, 2004). We also used the DNA curvature analysis website by Gohlke (https://www.lfd.uci.edu/~gohlke/dnacurve/) based on nucleosome positioning (Goodsell and Dickerson, 1994) to produce 3D models of the *Tyba* subfamilies sequences by opening the Helix Coordinate PDB file in the Geneious software.

RESULTS

Phylogenetic framework

The phylogenetic reconstruction of *Rhynchospora* based on nuclear sequences showed five main clades (**Fig. 1**). According to our dating analysis, the crown age of the genus was estimated at 37.8 Mya (95% CI = 33 – 42.1). Clade V was the first to diverge, with crown age of about 25.6 Mya (95% CI = 22.2 – 28.8). Later, clade I (which comprises species traditionally classified as part of *Eurhynchosporae*; Gale, 1944) diverged from the remaining clades (31.8 Mya, 95% CI = 28-35.2), with clade IV later diverging from the rest at around 30.9 Mya (95% CI = 26.6-34.5). The remaining species split into clade II and clade III (species formerly recognized as *Pleurostachys*, just recently synonymized as *Rhynchospora* section *Pleurostachys;* Thomas 2020) at 27.6 Mya (95% CI = 24.5-31.2).

Tyba is a diverse satDNA family with several variants

The number of analysed reads of our ISC analyses ranged from 173,544 reads in *R. glaziovii* to 2,719,919 in *R. pubera* **[Supplementary Table 1]**. General satDNA abundance varied from 0.02% in *R. emaciata* to 9.14% in one of the accessions of *R. corymbosa* **[Supplementary Table 3]**. From the 537 satDNA consensus sequences found in the individual clustering analysis of all species, 133 were found to have at least 60% sequence similarity with one of the previously published *Tyba* sequences, with monomer length varying from 170 to 175 bp. These *Tyba*-like sequences were found in most species of clades I-IV, with only two species (*R. biflora* and *R. racemosa*) of clade II showing no trace of *Tyba*. Satellite DNAs from species of clade V did not show significant similarity with *Tyba*. We further confirmed that these species from clades II and V did not present *Tyba* by being unable to find *Tyba*-like sequences in the full set of raw reads.

In our CC analysis, we found 23 clusters that were annotated by TAREAN as satellite DNAs **[Supplementary Table 2]**. From these, 17 clusters were found to have significant

amount of reads in more than one species, being considered shared satDNAs (**Fig. 2**). Six clusters of shared satDNAs (CL1, CL24, CL38, CL41, CL134 and CL141) were automatically annotated as having reads similar to *Tyba,* which was confirmed by the dotplot analysis **[Supplementary Figure 2]**. The consensus sequences of these satDNAs were found to have a 61.7% pairwise identity. Based on the high similarity and RE annotation, we determined that these satDNAs belong to the same satDNA family. An AML phylogeny showed that these six *Tyba*-like satDNAs could be divided in three groups based on sequence similarity (subfamilies A, B and C) (see **Supplementary Figure 1**). Along with *Tyba*, we also annotated two further shared satDNA clusters that were found to be of the same family **[Supplementary Figure 2]** and presented high sequence similarity with *RgSat*, a non-centromeric satDNA found by Ribeiro *et al.* (2017). The clusters identified as *RgSat* also presented moderate similarity with CL06 (48.4% pairwise identity) and with the *Tyba* A subfamily sequences (46.0%), but these values were below our threshold for being considered of the same family.

In the TLC analysis, the 276,052 input reads were divided into 12 different clusters, each assigned to one of the main *Tyba* subfamilies by similarity hits to the custom repeat database (see **Supplementary Figure 1, Supplementary Table 2**). An AML phylogenetic tree with the consensus sequence of the 12 *Tyba*-like clusters (TL-CL) showed that these could also be divided into different clades. Each of these clades were assigned to one of the *Tyba* subfamilies discovered in the shared satDNA analysis based on the RepeatExplorer annotation. Thus, each of the TL-CL were treated as one variant of a specific subfamily of *Tyba* **[Supplementary Figure 1]**.

Tyba variants are clade-specific within Rhynchospora

The CC analysis showed that *Tyba* was present among species of clades I, II, III and IV of *Rhynchospora*, with one of each Tyba subfamily being dominant in one specific clade, while subfamily B was present in both clades II and III (**Fig. 2**). None other satDNA was found in the CC analysis in more than one major clade. The *Rg*Sat satellite was found in clusters CL168 and CL332 (69.5% pairwise identity) and was present in only two species of clade V (*R. globosa* and *R. rubra*).

The TLC analysis mainly corroborated the results from the CC analysis (**Fig. 2**), with each clade presenting only one dominant *Tyba* subfamily (**Fig. 1**). Species of clades II and III showed only one variant from subfamily B per species. In contrast, in the clades I and IV, a few subclades presented more than one variant from the same subfamily (A and C, respectively, **Fig. 1**).

Tyba *sequences present high phylogenetic signal*

Although discrepancies between AAF tyba-like and hyb-seq topologies were observed at higher levels, we could retrieve a large number of monophyletic subclades congruent with the hyb-seq tree (**Fig. 3**). Most of the relationships inside clade I were well recovered in the AAF topology, with a few discrepancies, such as *R. capitellata* and *R. bucherorum* (**Fig. 3**). The only subclade from clade III that was not monophyletic in the AAF analysis was *R. tenuis* (*R. emaciata* + *R. riparia*). Support values of most nodes of the AAF tree were higher than 95%, with few lower supports at shallow levels and at the base of the clade containing species from clades II, III and IV (**Fig. 3**). In contrast, the AAF phylogeny based on the non-*Tyba* satDNAs had overall more incongruences when compared with the nuclear phylogeny **[Supplementary Figure 3]**. While clades I and III were recovered as monophyletic, species from the remaining clades were scattered throughout the phylogeny.

To investigate if *Tyba* sequence variability reflected phylogenetic relationships, we also assessed Pagel's λ for the GC content (%) and monomer size (bp) of the most abundant *Tyba* consensus sequence of each species. We found a high phylogenetic signal for both GC (λ = 0.92, $p < 0.001$) content and monomer size (λ = 0.93, $p < 0.001$), indicating that *Tyba* sequences from closely related species have similar nucleotide composition and monomer size, while distant species do not share similar values of these traits.

Tyba sequences are remarkably different from other satDNAs

The pairwise identity for all satDNA consensus mapped to each shared satDNA was calculated based on the resulting alignments and these identity values were plotted against the age of the most common recent ancestor (MRCA) of the species that shared that satDNA (**Fig. 4**). *Tyba* subfamilies presented lower pairwise identity than most of the other shared satDNAs (89.3%, 76.4% and 82.3% respectively for *Tyba* A, B and C), but were persistent through a longer evolutionary time (10, 28 and 27 Mya, respectively, for *Tyba* A, B and C). Overall, all consensus satDNA sequences from the ICS analysis assigned to the *Tyba* family preserved a pairwise identity of 70.9%, with the MRCA of the species that contained *Tyba* having originated 31 Mya. The other shared satDNAs had high pairwise identity values (from 82.8% on CL10 to 100% in CL370), but were only present on species that diversified more recently, from 2 Mya (CL370) to 8 Mya (CL61) (**Fig. 4**). The only non-*Tyba* satDNA that was preserved through a relatively long evolutionary time was *RgSat*. Although it was only presented in two species, those species shared a MRCA at around 21 Mya. However, the pairwise identity between the satDNAs mapped to the *RgSat* consensus was 69.1%. Overall, pairwise identity had a quicker decline through evolutionary time in non-*Tyba* satellites, while the *Tyba* subfamilies maintained

moderate pairwise identity values through relative longer periods, as the trend lines on **Fig. 4** indicate.

To check sequence conservation of the *Tyba* variants, the consensus sequences of the 12 variants were mapped to the consensus sequence of the three *Tyba* subfamilies, with identity values being calculated in a 2 bp slide window. Overall, *Tyba* A had a smaller number of mutations among its variants, with a high identity throughout most of the sequence (**Fig. 5**). Both *Tyba* B and *Tyba* C variants showed a larger number of regions bearing low identity, but also presented highly conserved motifs.

Curvature and bendability for the *Tyba* subfamilies and the other shared satDNAs were calculated based on DNase-I parameters. Curvature peaks were found in regions that were fairly conserved in the variants of subfamilies A and C, while it was not the case for the overall less conserved variants of subfamily B (**Fig. 5**). Similarly, bendability values were higher at conserved regions separated by a less conserved region between them, which presented the lowest bendability values (**Fig. 5**). The other shared satDNAs presented varied curvature/bendability profiles. Most sequences presented multiple curvature peaks, and bendability values did not follow the same pattern observed in *Tyba* **[Supplementary Figure 3]**. The *Rg*Sat sequence in contrast presented a single thin curvature peak and overall low bendability values, with only one peak as well **[Supplementary Figure 3]**.

DISCUSSION

The holocentromeric satDNA Tyba is conserved across Rhynchospora *phylogeny*

We demonstrated that the holocentromeric satellite DNA *Tyba* is evolutionary persistent and a relatively old family of satDNA in the genus *Rhynchospora*, especially when compared with the other shared satDNAs of the genus. Usually, satDNAs have a fast rate of evolution and just closely related species share satellite families (Lower *et al*., 2017). In more extreme cases, sequence divergence between centromeric satDNAs of different chromosomes in the same karyotypes have been reported (Zhang *et al*., 2014).

The library model for satDNA evolution postulates that, depending on the homogenization rates (under concerted evolution), it is possible to find a low number of copies of a satDNA family in a phylogenetically distant species, while that family can be dominant on genomes in other closely related species (Plohl *et al*., 2012; Garrido-Ramos, 2015). Next Generation Sequence methods have provided robust evidence for the library model of satDNA evolution, by facilitating the detection of the same satDNA families in phylogenetic unrelated species (Mravinac *et al*., 2002; Navajas-Pérez *et al*., 2009*b*; Quesada del Bosque *et al*., 2011). Here, *Tyba* was found in four out of the five main clades of *Rhynchospora* and its monomer sequence could be classified into three subfamilies, with each clade presenting a single type. We could find different variants of *Tyba* in different abundances across closely related species, especially in clades I and IV. Although many species presented only one *Tyba* variant, it is important to note that we used reads in a way comparable to a genome skimming analysis (Costa *et al*. 2021). This means that we had a small coverage for most of the analysed species, which in turn could mask low-abundant *Tyba* variants in the RepeatExporer analysis (Novák *et al*., 2017).

The propelling mechanisms of satDNA Tyba *evolution*

Different mechanisms seem to explain the diversification of *Tyba* holocentromeric satDNA in *Rhynchospora*. The high phylogenetic signal of GC content and monomer length, as well as the moderate congruence between our *Tyba*-like AAF phylogeny and the hyb-seq phylogeny are evidence of efficient clade/subclade-specific sequence homogenization of *Tyba*. Although the AAF phylogeny and the phylogenetic signal analysis used different types of data (raw NGS reads and TAREAN consensus sequences, respectively), both analyses showed that closely related species have more similar sequences of *Tyba* than distant species. Concerted evolution can lead to fast divergence between the satellites of reproductively isolated organisms, making these sequences potentially informative phylogenetic markers (Kuhn *et al*., 2012; Lorite *et al*., 2017; Belyayev *et al*., 2019).

Our results suggest that both concerted evolution and the library model explain *Tyba* evolution in *Rhynchospora*. It has been demonstrated that mechanisms of concerted evolution can led to the homogenization of different satDNA subfamilies that may have composed a genomic library of a group of species (Plohl *et al*., 2010; Ahmad *et al*., 2020). By this view, all three subfamilies of *Tyba* might have existed in the satDNA library of the ancestor of all species that possess *Tyba*, as it is speculated for other phylogenetic conserved satDNA families (Quesada del Bosque *et al*., 2011, 2014; Lorite *et al*., 2017). However, approx. 30 My of evolutionary time may have led to the observed clade-specific homogenization. One of the results of this clade-specific homogenization is that the clade-specific homogenized sequences may become useful phylogenetic markers (Kuhn *et al*., 2012; Lorite *et al*., 2017; Belyayev *et al*., 2019). Although homogenization can occur independently from natural selection, it is possible that favourable characteristics could led to evolutionary success of individuals presenting specific centromeric sequences, which could led to the selection for particular sequence variants within a centromeric satDNA array (Pérez-Gutiérrez *et al*., 2012; Lower *et al*., 2018).

Upon its discovery, two variants of *Tyba* were identified in *R. pubera* (Marques *et al*., 2015) and only one of these two in the closely related *R. tenuis* (Ribeiro *et al*., 2017), both positioned in clade IV. In our analysis, we could observe a total of 12 different variants within three subfamilies in each of the major *Rhynchospora* clade. Thus, patterns of both library model and concerted evolution could be found in clade IV. While younger lineages present up to three different *Tyba* variants in its composition, older lineages have only one dominant variant, which is congruent with the idea that concerted evolution is more efficient at long evolutionary times (Pérez-Gutiérrez *et al*., 2012). However, it could also be that the lack of other variants (even in small abundances) in these older lineages might indicate that speciation events which led to the more recent lineages might have facilitated the rise of new variants.

We could find both variants previously discovered in *R. pubera* (Marques *et al*., 2015) and a third, new variant in this species. Comparatively, *Tyba* sequences from species of clades II and III appear to have had a faster homogenization process, with only a single variant being dominant in each species, but the same variant shared with other species of the clade or even between clades, confirming its older origin. This could lead to clade III and a subclade of clade II appearing as monophyletic in the AAF topology. A number of factors have been proposed to explain why homogenization rates may vary between satDNA sequences, such as array length, organization and genomic location (Navajas-Pérez *et al*., 2009*a*; Kuhn *et al*., 2012; Lorite *et al*., 2017). The two variants of the oldest non-*Tyba* shared satDNA found in our CC analysis, *Rg*Sat, showed only 69% pairwise simmilarity. This satellite, first found in *R. globosa*, presents a traditional block-like distribution at the terminal regions of metaphase chromosomes and a clustered disposition at the chromocenters of interphasic nuclei (Ribeiro *et al*. 2017). *Tyba*, on the other hand, has been cytologically confirmed to be co-localized with the holocentromeres of species from clades II and IV (Marques *et al.* 2015; Rocha *et al*. 2016; Ribeiro *et al.* 2017; Costa *et al*. 2021) and has shown a dispersed distribution in interphasic nuclei (Marques *et al*.,

2015; Ribeiro *et al.* 2017). Thus, the fact that a dispersed satDNA present such high rates of homogenization, even when compared with a traditional block-like satDNA, is remarkable.

The *Eurhynchosporae* clade (formerly a tribe; Gale, 1944), corresponding to clade I in our phylogeny, presents an interesting case for satDNA evolution in a fast-diversifying group. This clade, which here appears as the first lineage to diverge among the clades that contain *Tyba*, was shown to present an increase in the diversification rate (Larridon *et al*. 2021). In situations of rapid diversification, sequence evolution is often not quick enough for sequences to diverge significantly between species (Giarla and Esselstyn, 2015). By this view, species within a recent diversification event tend to also present similar dominant satDNA variants (Quesada del Bosque *et al*., 2013; Dogan *et al*., 2021). However, while the A1 variant is widespread among clade I in our analyses, the other three *Tyba* A variants appear in specific subclades at comparable abundance, in a classic library model display.

The evolutionary persistence of Tyba *is related to its centromeric distribution*

Given its apparent phylogenetic conservation through long evolutionary times, what factors may be responsible for such long evolutionary reach of *Tyba*? It has been proposed by several studies that this long-time conservation of satDNA sequence may suggest structural/nucleotypic functional roles for these elements (Mravinac *et al*., 2005; Lower *et al*., 2018). Immuno-FISH experiments and ChIP-seq with the centromeric protein CENH3 have shown that *Tyba* is located at holocentromeres of four different *Rhynchospora* species (Marques *et al.* 2015; Rocha *et al*. 2016; Ribeiro *et al.* 2017; Costa *et al.* 2021), which could indicate functionality constraints. Although we cannot ascertain that every satDNA identified here as *Tyba*, or derivated from *Tyba*, has holocentromeric distribution, it may have a conserved chromosomal location because *Tyba* has been shown to be centromere-specific in phylogenetically distant species, such as *R. cephalotes* (Costa *et al*. 2021) and *R. pubera*

(Marques *et al*., 2015). Furthermore, the remarkably conserved sequence composition, monomer length and high abundance may be yet another indicator that all *Tyba* found here are indeed centromeric, seeing as centromeric satDNAs are often the most abundant in most organisms (Melters *et al*., 2013). Third, whole sequencing of *Rhynchospora* genomes demonstrated the centromeric role of *Tyba*, which can impact in evolutionary processes such as genomic architecture, epigenome organization, and karyotype evolution (Hofstatter *et al*. 2022).

Although satellite DNA is a major constituent of eukaryotic centromeres, sequence conservation is extremely low in most cases (Plohl *et al*., 2014; Garrido-Ramos, 2017). In fact, centromeric satDNAs have been shown to differ among chromosomes of the same karyotype in some plant species (Gong *et al*., 2012; Iwata *et al*., 2013; Zhang *et al*., 2014). An investigation of putative centromeric satDNA of 282 species across eukaryotes showed that sequence similarity rapidly declines through divergence, reaching background noise levels after 50 My of divergence (Melters *et al*., 2013), which would make *Tyba*, with 31 My, relatively old when compared with most centromeric satDNAs.

Our curvature/bendability analysis have shown the highest values of curvature and bendability at similar regions of the *Tyba* subfamilies sequences. In addition, this analysis also showed that *Tyba* sequences are more flexible than the sequence of *Rg*Sat, a confirmed non-centromeric satDNA. Other shared satDNAs presented variable patterns of curvature, with multiple peaks and mostly constant bendability values throughout the sequences. The curvature patterns of *Tyba* are similar to the proposed patterns of core DNA (tight sequences that wrap around a histone), in which large curvature at the ends of the sequence facilitate nucleosome formation (Liu *et al*., 2008). In the same way, high and stable bendability values for satellite sequences were also proposed to facilitate nucleosomal organization of centromeric proteins

(Escudero *et al*., 2019). Our data is supported by genomic analyses that confirmed the centromeric role of satDNA Tyba (Hofstatter *et al*. 2022).

One important observation is that *Tyba* is is not present in all *Rhynchospora* species, being absent in the whole clade V and in few species from clade II, which emphasizes that *Tyba* is not necessary for holocentricity. For example, the clade V species *R. globosa*, which has meiotic evidence of its holocentricity (Arguelho *et al*., 2012), presented only satDNAs disposed in block-like patterns (Ribeiro *et al*., 2017). While *Tyba* is probably not necessary for the holocentromere of *Rhynchospora* species, it is possible that its sequence confers advantages regarding nucleosome formation and positioning which may have resulted in its conservation through a long evolutionary time (Talbert and Henikoff, 2020).

It has also been showed that satDNAs can be associated with retrotransposons at functional parts of the centromere, as is the case with the satDNA CentO and retrotransposon CRR of rice (Cheng *et al*., 2002). A similar association was found in *R. pubera*, where the *Ty3-Gypsy* retroelement CRRh was shown to have a similar pattern as *Tyba*, spread throughout the holocentromeres (Marques *et al*., 2015). It has been proposed that an association with mobile elements could facilitate the spread of centromeric satDNAs. Most notably, the association with mobile elements seemed to have facilitate the spread and persistence of relic satDNA BIV160 for over 500 My of evolution (Plohl *et al*., 2010). Thus, the association with mobile elements could favour the conservation of *Tyba* sequences across *Rhynchospora* holocentric chromosomes, providing the mobility and means of fixation for this satDNA.

Here, 404 satDNA consensus sequences found in the individual species clustering analysis (mean of approx. 5 per accession) could not be mapped to one of the *Tyba* and non-*Tyba* shared satDNAs, meaning that those were probably species-specific. The oldest non-*Tyba* shared satDNA (*Rg*Sat, previously found in *R. globosa* by Ribeiro *et al.* 2017) was only present in two distantly related species (*R. rubra* and *R. globosa*) that shared a common ancestor around

21 My. By contrast, clades containing *Tyba* shared a 31.8My common ancestor and maintained a pairwise identity over 70%. While the concept of evolutionary old satDNAs may be uncommon, it is not exactly rare (see Plohl *et al.* 2010; Quesada del Bosque *et al.* 2013). The existence of these "relic" sequences may indicate that this type of repeat could provide evolutionary advantages by playing a functional role, leading to a long-time conservation (Plohl *et al*., 2012; Lower *et al*., 2018).

Here, we have unravelled the evolutionary history of the holocentromere-specific satellite *Tyba* (Marques *et al*., 2015, 2016; Ribeiro *et al*., 2018) in what is one of the largest samplings for satDNA-related analysis to date. We observed a high sequence conservation and phylogenetic significance for *Tyba*, especially when compared with other satDNAs observed in the genus. *Tyba* seems to have evolved by a combination of satDNA library diversification and concerted evolution. The sequence has shown to persist throughout a long evolutionary time, with clade-specific variants, and was present in most of the genus, in striking contrast to the other satDNAs found in *Rhynchospora.* Our results propose that such conservation could be linked to a functional role within the holocentromeres. While we present a thorough outline of *Tyba* evolution, future studies with long-read sequencing techniques may help to better understand the organization of *Tyba* arrays, which could provide further clues about its intimate association with the centromeres of *Rhynchospora*.

FUNDING INFORMATION

This study was supported in part by the Coordenacão de Aperfeicoamento de Pessoal de Nıvel Superior–Brasil (CAPES) (Finance Code 001), CAPES-PRINT (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Programa Institucional de Internacionalização) [project number 88887.363884/2019-00 (L.C.)] and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnologico) [grant number 141037/2018-0 (L.C.)].

ACKNOWLEDGEMENTS

We thank Dr. Diogo Cabral-de-Melo (Universidade Estadual de São Paulo), Dr. André Luís Laforga Vanzela (Universidade Estadual de Londrina), Dr. Giovana Augusta Torres (Universidade Estadual de Lavras) and Dr. Lyderson Facio Viccini (Universidade Federal de Juiz de Fora) for providing comments and suggestions for the manuscript.

LITERATURE CITED

Achrem M, Szućko I, Kalinka A. 2020. The epigenetic regulation of centromeres and telomeres in plants and animals. *CCG* 14: 265–311. doi:10.3897/CompCytogen.v14i2.51895.

Ahmad SF, Singchat W, Jehangir M, *et al*. 2020. Dark Matter of Primate Genomes: Satellite DNA Repeats and Their Evolutionary Dynamics. *Cells* 9: 2714. doi:10.3390/cells9122714.

Arguelho EG, Michelan VS, Nogueira FM, *et al*. 2012. New chromosome counts in Brazilian species of *Rhynchospora* (Cyperaceae). *Caryologia* 65, 140–146. doi:10.1080/00087114.2012.711675.

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. *IJMS* 20: 1201. doi:10.3390/ijms20051201.

Brukner I, Sánchez R, Suck D, Pongor S. 1995. Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *EMBO J* 14: 1812–1818.

Buddenhagen CE. 2016. *A view of Rhynchosporeae (Cyperaceae) diversification before and after the application of anchored phylogenomics across the angiosperms*. PhD Thesis, Florida State University, USA.

Burchardt P, Buddenhagen CE, Gaeta ML, Souza MD, Marques A, Vanzela ALL. 2020. Holocentric Karyotype Evolution in *Rhynchospora* Is Marked by Intense Numerical, Structural, and Genome Size Changes. *Frontiers in Plant Science* 11: 536507. doi:10.3389/fpls.2020.536507.

Bureš P, Zedek F, Markova M. 2013. Holocentric Chromosomes. In: Greilhuber J, Dolezel J, Wendel JF, eds. *Plant Genome Diversity Volume 2*. Vienna: Springer, 187–204.

Camacho JPM, Cabrero J, López-León MD *et al*. (2021). Satellitome comparison of two oedipodine grasshoppers highlights the contingent nature of satellite DNA evolution. *BMC Biology*, 20: 36. doi:10.1101/2021.07.01.450629.

Cheng Z, Dong F, Langdon T *et al*. 2002. Functional Rice Centromeres Are Marked by a Satellite Repeat and a Centromere-Specific Retrotransposon. *The Plant Cell* 14: 1691–1704. doi:10.1105/tpc.003079.

Costa L, Marques A, Buddenhagen C, *et al*. 2021. Aiming off the target: recycling target capture sequencing reads for investigating repetitive DNA. *Annals of Botany*, 128: 835-848. doi:10.1093/aob/mcab063.

Dogan M, Pouch M, Mandáková T, *et al*. 2021. Evolution of Tandem Repeats Is Mirroring Post-polyploid Cladogenesis in *Heliophila* (Brassicaceae). *Frontiers in Plant Science* 11: 607893. doi:10.3389/fpls.2020.607893.

Dover G. 2002. Molecular drive. *Trends in Genetics* 18: 587–589. doi:10.1016/S0168-9525(02)02789-0.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* 7: 214. doi:10.1186/1471-2148-7-214.

Escudeiro A, Adega F, Robinson TJ, Heslop-Harrison JS, Chaves R. 2019. Conservation, Divergence, and Functions of Centromeric Satellite DNA Families in the Bovidae. *Genome Biology and Evolution* 11: 1152–1165. doi:10.1093/gbe/evz061.

Fan H, Ives AR, Surget-Groba Y, Cannon CH. 2015. An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. *BMC Genomics* 16: 522. doi:10.1186/s12864-015-1647-5.

Fry K, Salser W. 1977. Nucleotide sequences of HS-α satellite DNA from kangaroo rat dipodomys ordii and characterization of similar sequences in other rodents. *Cell* 12: 1069–1084. doi:10.1016/0092-8674(77)90170-2.

Gabrielan E, Zohary D. 2004. Wild relatives of food crops native to Armenian and Nakhichevan. *Flora Mediterranea* 14: 5–80.

Gale S. 1944. *Rhynchospora*, section *Eurhynchospora*, in Canada, the United States and the West Indies. In Gale S, eds. *Contributions from the Gray Herbarium of Harvard University*, New England Botanical Club, 89–134

Garrido-Ramos M. 2017. Satellite DNA: An Evolving Topic. *Genes* 8: 230. doi:10.3390/genes8090230.

Garrido-Ramos MA. 2015. Satellite DNA in Plants: More than Just Rubbish. *Cytogenetic and Genome Research* 146: 153–170. doi:10.1159/000437008.

Gemmell NJ. 2021. Repetitive DNA: genomic dark matter matters. *Nature Reviews Genetics* 22: 342–342. doi:10.1038/s41576-021-00354-8.

Gernhard T. 2008. New Analytic Results for Speciation Times in Neutral Models. *Bulletin of Mathematical Biology* 70: 1082–1097. doi:10.1007/s11538-007-9291-0.

Giarla TC, Esselstyn JA. 2015. The Challenges of Resolving a Rapid, Recent Radiation: Empirical and Simulated Phylogenomics of Philippine Shrews. *Systematic Biology* 64: 727–740. doi:10.1093/sysbio/syv029.

Gong Z, Wu Y, Koblížková A, *et al*. 2012. Repeatless and Repeat-Based Centromeres in Potato: Implications for Centromere Evolution. *The Plant Cell* 24: 3559–3574. doi:10.1105/tpc.112.100511.

Goodsell DS, Dickerson RE. 1994. Bending and curvature calculations in B-DNA. *Nucleic Acids Research* 22: 5497–5503. doi:10.1093/nar/22.24.5497.

Heckmann S, Schroeder-Reiter E, Kumke K, *et al*. 2011. Holocentric Chromosomes of *Luzula elegans* Are Characterized by a Longitudinal Centromere Groove, Chromosome Bending, and a Terminal Nucleolus Organizer Region. *Cytogenetic and Genome Research* 134: 220–228. doi:10.1159/000327713.

Hofstatter PG, Thangavel G, Lux T, *et al.* 2022. Repeat-based holocentromeres influence genome architecture and karyotype evolution. *Cell* 185: 3153-3168.

Larridon I, Spalink D, Jiménez-Mejías P, *et al.* 2021. The evolutionary history of sedges (Cyperaceae). *Madagascar. Journal of Biogeography* 48: 917-932. doi: 10.1111/jbi.14048

Iwata A, Tek AL, Richard MMS, *et al*. 2013. Identification and characterization of functional centromeres of the common bean. *Plant Journal* 76: 47-60. doi:10.1111/tpj.12269.

Kearse M, Moir R, Wilson A, *et al*. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28: 1647–1649. doi:10.1093/bioinformatics/bts199.

Kuhn GCS, Küttler H, Moreira-Filho O, Heslop-Harrison JS. 2012. The 1.688 Repetitive DNA of *Drosophila*: Concerted Evolution at Different Genomic Scales and Association with Genes. *Molecular Biology and Evolution* 29: 7–11. doi:10.1093/molbev/msr173.

Kükenthal G. 1939. Vorarbeiten zu einer Monographie der Rhynchosporoideae. *Feddes Repert* 47: 209–216. doi:10.1002/fedr.19390471502.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9: 357–359. doi:10.1038/nmeth.1923.

Liu H, Duan X, Yu S, Sun X. 2011. Analysis of nucleosome positioning determined by DNA helix curvature in the human genome. *BMC Genomics* 12: 72. doi:10.1186/1471-2164-12-72.

Liu H, Wu J, Xie J, Yang X, Lu Z, Sun X. 2008. Characteristics of Nucleosome Core DNA and Their Applications in Predicting Nucleosome Positions. *Biophysical Journal* 94: 4597–4604. doi:10.1529/biophysj.107.117028.

Lorite P, Muñoz-López M, Carrillo JA, *et al*. 2017. Concerted evolution, a slow process for ant satellite DNA: study of the satellite DNA in the *Aphaenogaster* genus

(Hymenoptera, Formicidae). *Organisms Diversity and Evolution* 17: 595–606. doi:10.1007/s13127-017-0333-7.

Lower SS, Johnston JS, Stanger-Hall KF, *et al*. 2017. Genome Size in North American Fireflies: Substantial Variation Likely Driven by Neutral Processes. *Genome biology and evolution* 9: 1499–1512. doi:10.1093/gbe/evx097.

Lower SS, McGurk MP, Clark AG, Barbash DA. 2018. Satellite DNA evolution: old ideas, new approaches. *Current Opinion in Genetics & Development* 49: 70–78. doi:10.1016/j.gde.2018.03.003.

Macas J, Meszaros T, Nouzova M. 2002. PlantSat: a specialized database for plant satellite repeats. *Bioinformatics* 18: 28–35. doi:10.1093/bioinformatics/18.1.28.

Marques A, Ribeiro T, Neumann P, *et al*. 2015. Holocentromeres in *Rhynchospora* are associated with genome-wide centromere-specific repeat arrays interspersed among euchromatin. *Proceedings of the National Academy of Sciences* 112: 13633–13638. doi:10.1073/pnas.1512255112.

Marques A, Pedrosa-Harand A. 2016. Holocentromere identity: from the typical mitotic linear structure to the great plasticity of meiotic holocentromeres. *Chromosoma* 125: 669–681. doi:10.1007/s00412-016-0612-7.

Marques A, Schubert V, Houben A, Pedrosa-Harand A. 2016. Restructuring of Holocentric Centromeres During Meiosis in the Plant *Rhynchospora pubera*. *Genetics* 204: 555–568. doi:10.1534/genetics.116.191213.

Melters DP, Bradnam KR, Young HA, *et al*. 2013. Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biology* 14: R10. doi:10.1186/gb-2013-14-1-r10.

Mravinac B, Plohl M, Mestrović N, Ugarković Đ. 2002. Sequence of PRAT Satellite DNA ``Frozen'' in Some Coleopteran Species. *Journal of Molecular Evolution* 54: 774–783. doi:10.1007/s00239-001-0079-9.

Mravinac B, Plohl M, Ugarković Đ. 2005. Preservation and High Sequence Conservation of Satellite DNAs Suggest Functional Constraints. *Journal of Molecular Evolution* 61: 542–550. doi:10.1007/s00239-004-0342-y.

Navajas-Pérez R, Quesada del Bosque ME, Garrido-Ramos MA. 2009*a*. Effect of location, organization, and repeat-copy number in satellite-DNA evolution. *Molecular Genetics and Genomics* 282: 395–406. doi:10.1007/s00438-009-0472-4.

Navajas-Pérez R, Schwarzacher T, Ruiz Rejón M, Garrido-Ramos MA. 2009*b*. Characterization of RUSI, a telomere-associated satellite DNA, in the genus *Rumex* (Polygonaceae). *Cytogenetic and Genome Research* 124: 81–89. doi:10.1159/000200091.

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 Research* 45: e111–e111. doi:10.1093/nar/gkx257.

Novak 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: 792–793. doi:10.1093/bioinformatics/btt054.

Oliveira MAS, Nunes T, Dos Santos MA *et al.* 2021. High-Throughput Genomic Data Reveal Complex Phylogenetic Relationships in *Stylosanthes* Sw (Leguminosae). *Frontiers in genetics* 12: 727314. doi: 10.3389/fgene.2021.727314

Pagel M. 1999. Inferring the historical patterns of biological evolution. *Nature* 401: 877–884. doi:10.1038/44766.

Pérez-Gutiérrez MA, Suárez-Santiago VN, López-Flores I, Romero AT, Garrido-Ramos MA. 2012. Concerted evolution of satellite DNA in *Sarcocapnos*: a matter of time. *Plant Molecular Biology* 78: 19–29. doi:10.1007/s11103-011-9848-z.

Petraccioli A, Odierna G, Capriglione T, *et al*. 2015. A novel satellite DNA isolated in *Pecten jacobaeus* shows high sequence similarity among molluscs. *Molecular Genetics and Genomics* 290: 1717–1725. doi:10.1007/s00438-015-1036-4.

Plohl M, Meštrović N, Mravinac B. 2014. Centromere identity from the DNA point of view. *Chromosoma* 123: 313–325. doi:10.1007/s00412-014-0462-0.

Plohl M, Meštrović N, Mravinac B. 2012. Satellite DNA Evolution. In: Garrido-Ramos MA, eds. *Genome Dynamics*, Basel: S KARGER AG, 126–152. doi:10.1159/000337122.

Plohl M, Petrović V, Luchetti A *et al*. (2010). Long-term conservation vs high sequence divergence: the case of an extraordinarily old satellite DNA in bivalve mollusks. *Heredity* 104, 543–551. doi:10.1038/hdy.2009.141.

Quesada del Bosque ME, López-Flores I, Suárez-Santiago VN, Garrido-Ramos MA. 2013. Differential spreading of HinfI satellite DNA variants during radiation in Centaureinae. *Annals of Botany* 112: 1793–1802. doi:10.1093/aob/mct233.

Quesada del Bosque ME, Navajas-Pérez R, Panero JL, Fernández-González A, Garrido-Ramos MA. 2011. A satellite DNA evolutionary analysis in the North American endemic dioecious plant *Rumex hastatulus* (Polygonaceae). *Genome* 54: 253–260. doi:10.1139/g10-115.

Quesada del Bosque MEQ, López-Flores I, Suárez-Santiago VN, Garrido-Ramos MA. 2014. Satellite-DNA diversification and the evolution of major lineages in *Cardueae* (Carduoideae Asteraceae). *Journal of Plant Research* 127: 575–583. doi:10.1007/s10265-014-0648-9.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing Available at: https://www.R-project.org/.

Rambaut A, Drummond AJ. 2013. *TreeAnnotator v1. 7.0. Available as Part of the BEAST package*. Available at: http://beast.bio.ed.ac.uk.

Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* 3: 217–223. doi:10.1111/j.2041-210X.2011.00169.x.

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: 263–272. doi:10.1007/s00709-017-1154-4.

Ribeiro T, Marques A, Novák P, *et al*. 2017. Centromeric and non-centromeric satellite DNA organisation differs in holocentric *Rhynchospora* species. *Chromosoma* 126: 325–335. doi:10.1007/s00412-016-0616-3.

Ribeiro T, Vasconcelos E, dos Santos KGB, Vaio M, Brasileiro-Vidal AC, Pedrosa-Harand A. 2020. Diversity of repetitive sequences within compact genomes of *Phaseolus* L. beans and allied genera *Cajanus* L. and *Vigna* Savi. *Chromosome Research* 28: 139–153. doi:10.1007/s10577-019-09618-w.

Robles F, de la Herrán R, Ludwig A, Ruiz Rejón C, Ruiz Rejón M, Garrido-Ramos MA. 2004. Evolution of ancient satellite DNAs in sturgeon genomes. *Gene* 338: 133–142. doi:10.1016/j.gene.2004.06.001.

Rocha DM, Marques A, Andrade CG, *et al*. 2016. Developmental programmed cell death during asymmetric microsporogenesis in holocentric species of Rhynchospora (Cyperaceae). *Journal of experimental botany* 67: 5391-5401.

Salser W, Bowen S, Browne D, *et al*. 1976. Investigation of the organization of mammalian chromosomes at the DNA sequence level. *Federation Proceedings* 35: 23–35.

Silva Filho PJS, Thomas WW, Boldrini II. 2021. Redefining *Rhynchospora* section *Tenues* (Cyperaceae), a phylogenetic approach. *Botanical Journal of the Linnean Society* boab002. doi:10.1093/botlinnean/boab002.

Smith SY, Collinson ME, Rudall PJ, Simpson DA. 2010. Cretaceous and Paleogene Foss il Record of Poales: Review and Current Research. In: Fay MF, eds. *Diversity, phylogeny, and evolution in the monocotyledons.* Denmark: Aarhus University Press, 335–356.

Smith SY, Collinson ME, Simpson DA, Rudall PJ, Marone F, Stampanoni M. 2009. Elucidating the affinities and habitat of ancient, widespread Cyperaceae: *Volkeria messelensis* gen. et sp. nov., a fossil mapanioid sedge from the Eocene of Europe. *American Journal of Botany* 96: 1506–1518. doi:10.3732/ajb.0800427.

Sonnhammer ELL, Durbin R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* 167: GC1–GC10. doi:10.1016/0378-1119(95)00714-8.

Subirana JA, Messeguer X. 2013. A Satellite Explosion in the Genome of Holocentric Nematodes. *PLoS ONE* 8: e62221. doi:10.1371/journal.pone.0062221.

Talbert PB, Henikoff S. 2020. What makes a centromere? *Experimental Cell Research* 389: 111895.

Thomas WW. 2020. Two new species of *Rhynchospora* (Cyperaceae) from Bahia, Brazil, and new combinations in Rhynchospora section Pleurostachys. *Brittonia* 72: 273–281. doi:10.1007/s12228-020-09621-0.

Tsoumani KT, Drosopoulou E, Mavragani-Tsipidou P, Mathiopoulos KD. 2013. Molecular Characterization and Chromosomal Distribution of a Species-Specific Transcribed Centromeric Satellite Repeat from the Olive Fruit Fly, *Bactrocera oleae*. *PLoS ONE* 8: e79393. doi:10.1371/journal.pone.0079393.

Zhang H, Koblížková A, Wang K *et al*. 2014. Boom-Bust Turnovers of Megabase-Sized Centromeric DNA in *Solanum* Species: Rapid Evolution of DNA Sequences Associated with Centromeres. *The Plant Cell* 26: 1436–1447. doi:10.1105/tpc.114.123877.

Zhang T, Talbert PB, Zhang W, *et al*. 2013. The CentO satellite confers translational and rotational phasing on cenH3 nucleosomes in rice centromeres. *Proceedings of the National Academy of Sciences* 110: E4875–E4883. doi:10.1073/pnas.1319548110.

Zhong CX, Marshall JB, Topp C, *et al*. 2002. Centromeric Retroelements and Satellites Interact with Maize Kinetochore Protein CENH3. *The Plant Cell* 14: 2825–2836. doi:10.1105/tpc.006106.

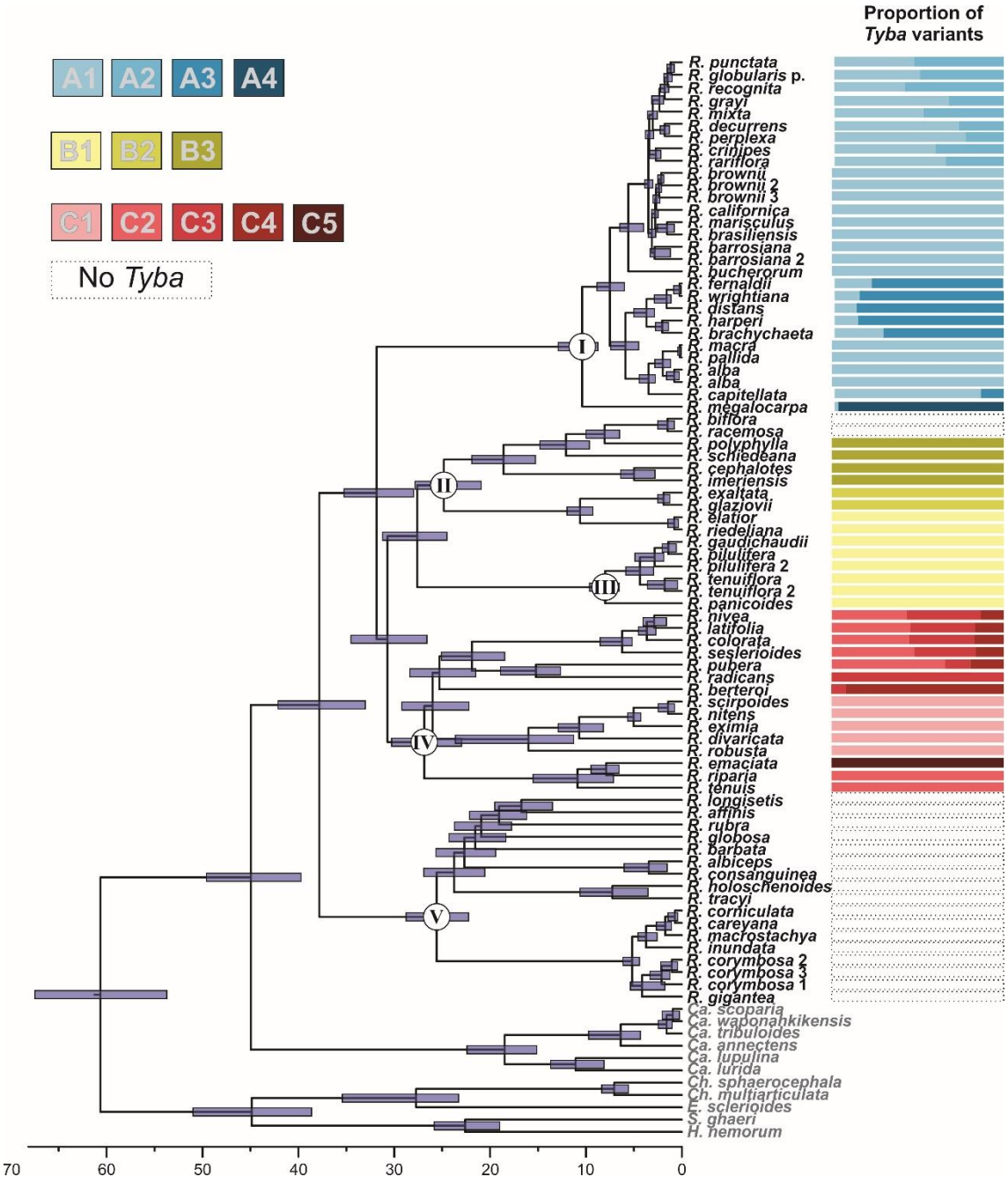


Figure 1 – Phylogenetic relationships and molecular dating of Rhynchospora species. Colored bars (according to the caption on the upper left of the figure) represent the proportion of different Tyba variants found in the genome of each species. Bars at the nodes of the tree represent the 95% confidence interval of the molecular dating analysis, with axis scale representing divergence time in millions of years. Circles with roman numbers in the nodes delimits the five clades discussed.

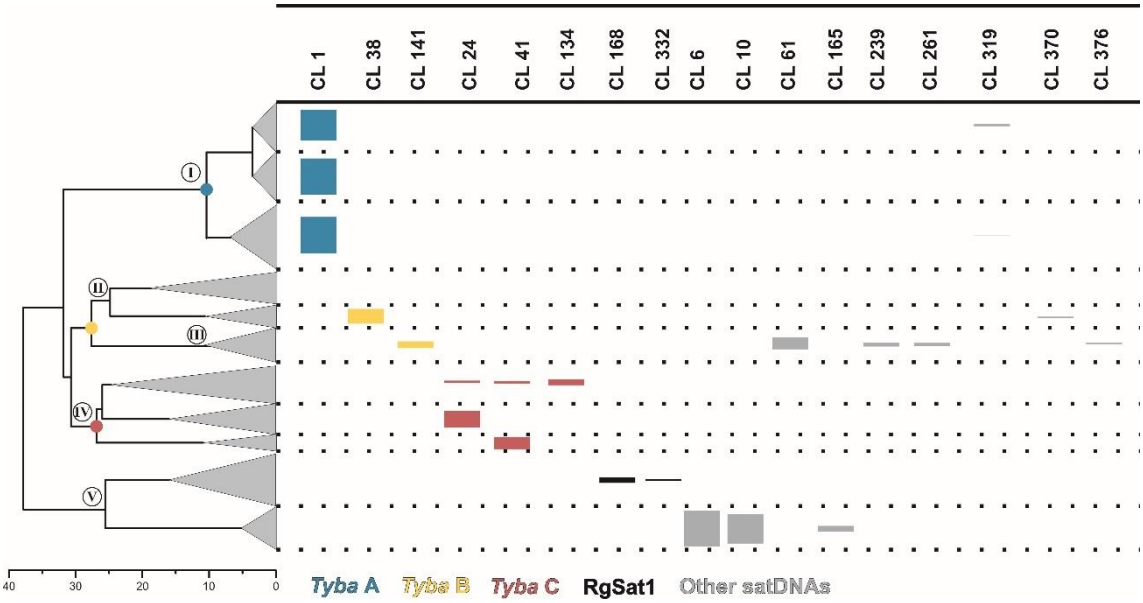


Figure 2 – Shared satellite DNAs in analyzed Rhynchospora. Each line represents one of the clades of the simplified Rhynchospora phylogeny on the left (based on the tree presented in Figure 1). Clusters annotated as satDNA by RepeatExplorer are presented in the upper part and their relative abundance in a given species is represented by the rectangle's height. Rectangle colors are according to the caption below. Small colored circles at the nodes of the simplified tree represent possible points of origin of the different Tyba subfamilies.

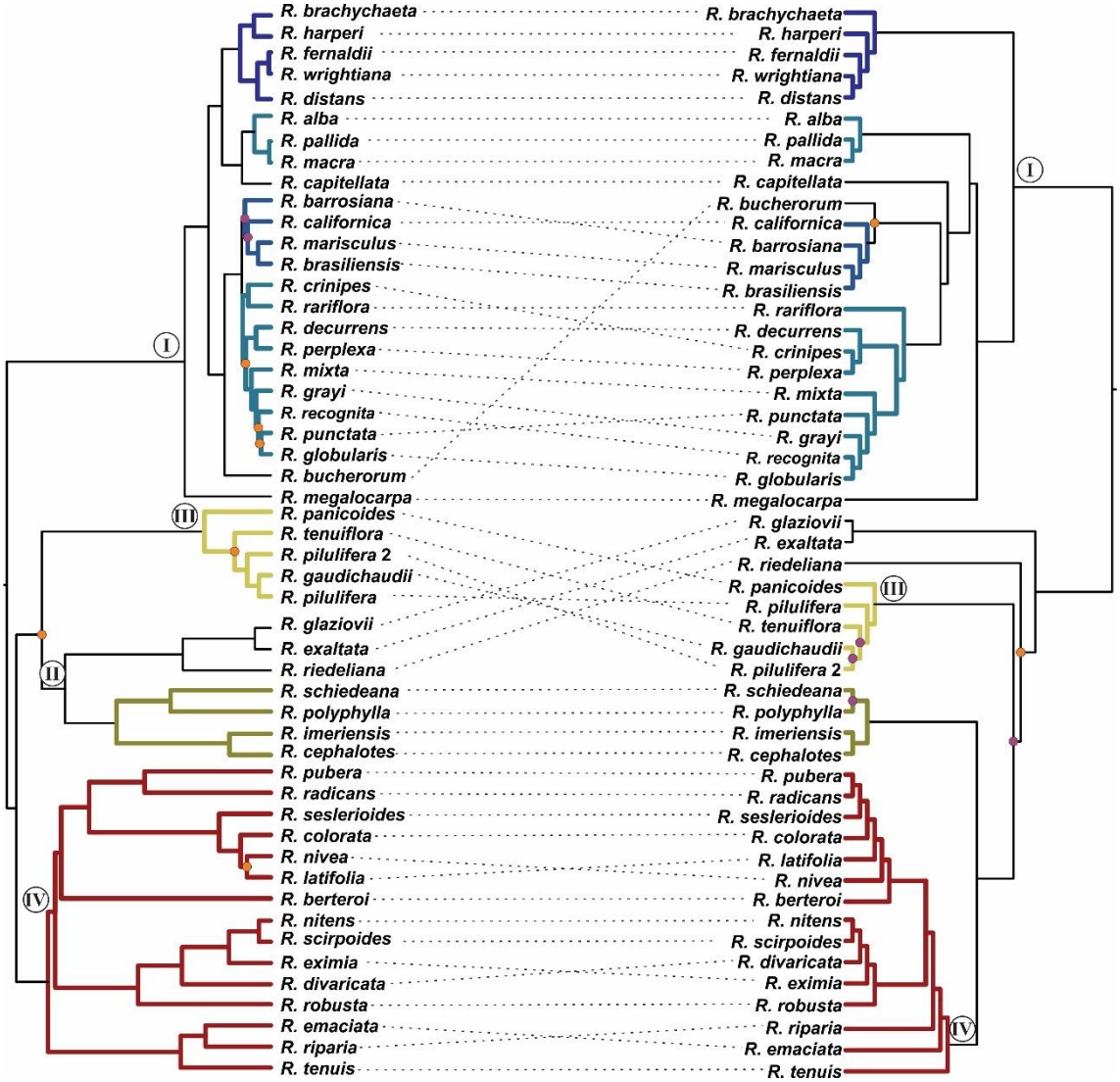


Figure 3 – Comparison of phylogenetic trees recovered from nuclear markers (left) and AAF of Tyba-like reads (right). Clades colored in shades of blue (clade I), yellows (clade II) and red (clade III) are monophyletic in both trees. Circles at tree nodes refer to support values between 75-95% (orange) and lower than 75% (purple). Nodes without circles presented maximum support.

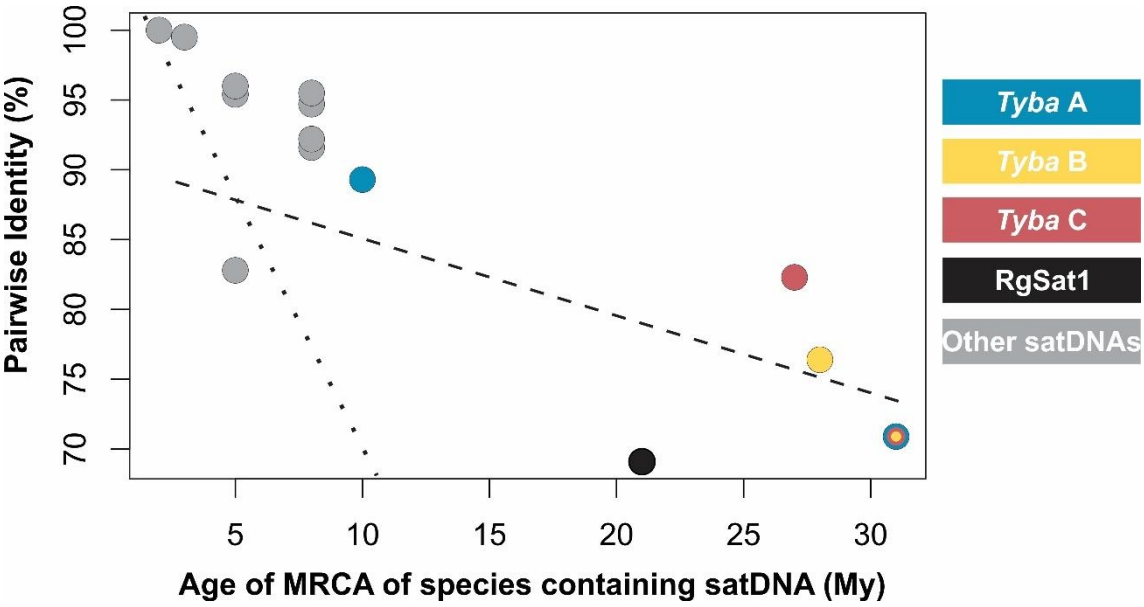


Figure 4 – Plot of pairwise identity through evolutionary time for the shared satDNAs recovered in the comparative clustering analysis. Dots represent the shared satDNAs, with colors reflecting the caption at the right side. Dotted trend line refers to non-Tyba satDNAs, while dashed trend line refers to the Tyba subfamilies. Multi-colored dot refers to the whole Tyba family.

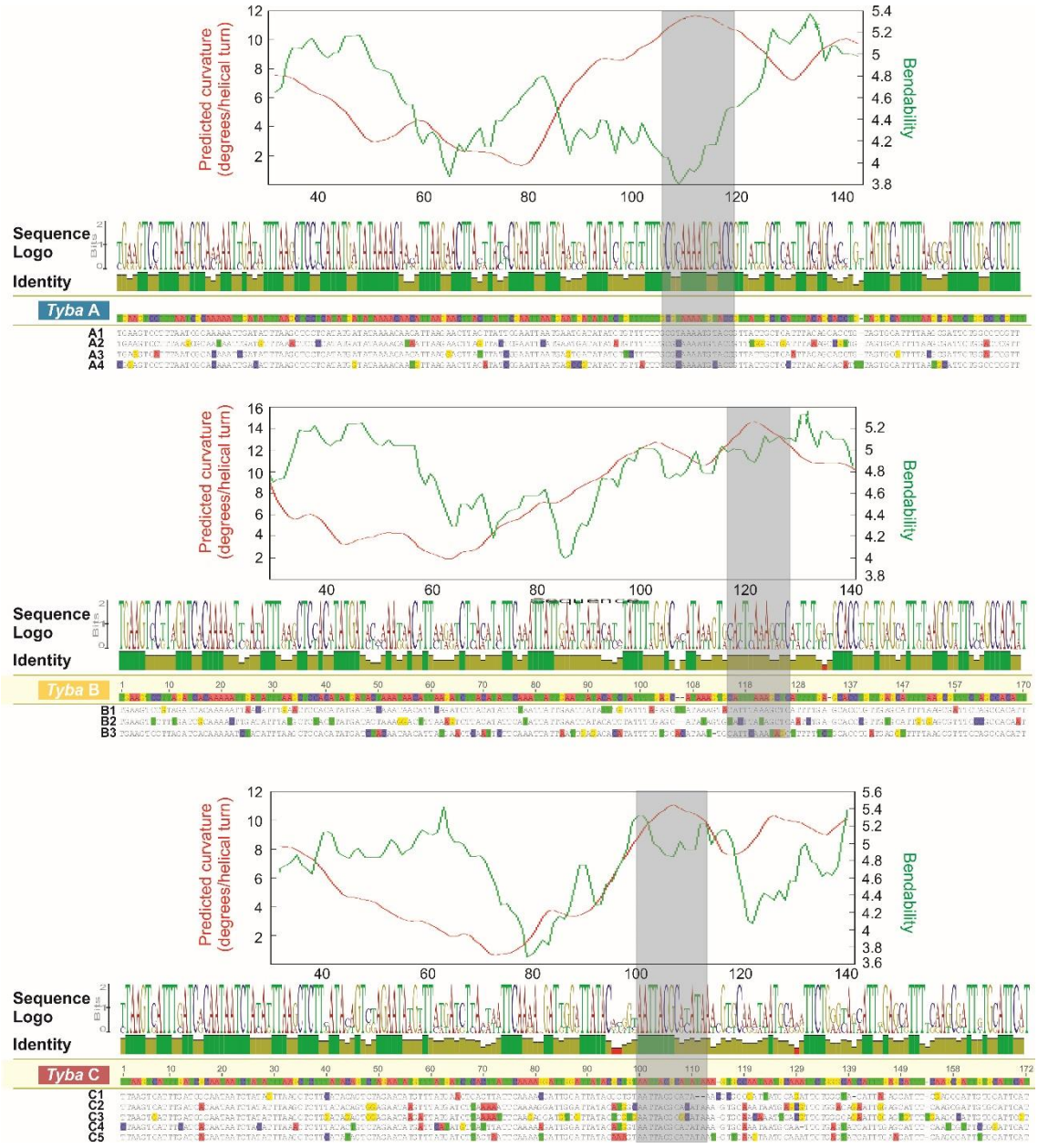


Figure 5 – Bendability (green)/Curvature (red) propensity, sequence logos, identity plot and alignment of Tyba variants assigned to each Tyba subfamily. Heights of nucleotides in the sequence logos are proportional to the number of times that base appeared in the alignment. Green values on the identity plot represent high sequence identity, while yellow represent moderate and red represent low sequence identity. Base pairs that are different from the consensus sequence are colored in the alignment. Gray rectangles represent the approximate region where the curvature peak was estimated.

Supplementary Table 1 – List of Rhynchospora species, voucher number, herbarium, number of reads before and after target-reads filtering and total number of reads analyzed in the individual clustering analysis into RepeatExplorer.

Species	Voucher – herbarium	Reads before filtering	Reads after filtering	Analyzed reads
Carex annectens (E.P.Bicknell) P.Bicknell	Andrew Hipp 2013014 – MOR	-	-	-
Carex lupulina Muhl. ex Willd.	Andrew Hipp 2013012 – MOR	-	-	-
Carex lúrida Wahlenb.	Loran C. Anderson 24918 – FSU	-	-	-
Carex scoparia Willd.	Marilee Lovit 464 – MOR	-	-	-
Carex tribuloides Wahlenb.	Andrew Hipp 2013001 – MOR	-	-	-
Carex waponahkikensis M.Lovit & A.Haines	Marilee Lovit 420B - MOR	-	-	-
Chorizandra multiarticulata Nees	Chrissie J. Prychid 40 – NE	-	-	-
Chorizandra sphaerocephala R.Br.	Chrissie J. Prychid 42 – NE	-	-	-
Exocarya sclerioides (F. Muell.) Benth	Chrissie J. Prychid 41 – NE	-	-	-
Hypolytrum nemorum (Vahl)	Chrissie J. Prychid 45 – NE	-	-	-
R. affinis W.Fitzg.	A.W. Scott 79225 – NE	4,451,598	4,177,006	1,516,574
R. alba (L.) Vahl	Bob Moyer – FSU	5,075,306	4,085,050	747,026
	Jeremy Bruhl 3088 – NE	5,566,262	4,588,986	674,321
R. albiceps Kunth	Wayt Thomas 16397 – NY	4,399,366	3,765,890	810,371
R. barbata (Vahl) Kunth	Wayt Thomas 16195 – NY	4,164,678	3,737,976	417,260
R. barrosiana Guagl.	Jeremy Bruhl 2330 – NE	5,127,108	4,177,638	762,091
	K.L. Wilson 10726 – NSW	4,113,776	3,482,476	938,696
R. berteroi (Spreng.) C.B.Clarke	Wayt Thomas 14864 - NY	5,127,130	4,370,998	978,959
R. biflora Boeckeler	Wayt Thomas 15002 – NY	3,621,336	3,228,896	1,104,253
R. brachychaeta Wright ex Sauv.	Orzell and Bridges 26718 – FSU	5,111,046	4,277,254	814,765
R. brasiliensis Boeckeler	Wayt Thomas 16402 – NY	4,377,466	3,579,108	623,699
R. brownii Roem. & Schult.	L.M. Copeland 3336 – NE	6,415,422	5,280,206	826,435
	Jeremy Bruhl 1714 – NE	7,049,162	5,247,494	768,855
	Jeremy Bruhl 2750 – NE	4,783,888	3,935,096	768,475
R. bucherorum León	Wayt Thomas 14933 – NY	4,492,178	3,831,184	1131705
R. californica Gale	Robert Naczi 10626 – DOV	5,251,328	4,312,636	621,597
R. capitellata (Michx.) Vahl	Bob Moyer s.n. – FSU	5,333,370	4,387,888	794,488
R. careyana (Michx.) Vahl	Chris Buddenhagen 1309072 – FSU	3,758,370	3,012,434	465,498
R. cephalotes (L.). Vahl	Robert Naczi 11691 – NY	6,932,932	5,330,038	580,881
R. colorata (L.) H. Pfeiffer	Chris Buddenhagen 1307091 – FSU	4,660,462	3,780,928	866,372
R. consanguinea (Kunth) Boeckeler	Wayt Thomas 16404 – NY	4,505,686	3,887,754	653,574
R. corniculata (Lam.) A.Gray	Loran C. Anderson 27381 - FSU	5,654,854	4,928,976	688,746
R. corymbosa (L.) Britton	Jeremy Bruhl 2316 – NE	3,486,332	3,033,422	221,871
	Jeremy Bruhl 2521 – NE	4,290,778	3,674,188	425,608
	Jeremy Bruhl 3085 – NE	2,599,950	2,237,996	261,444
R. crinipes Gale	Chris Buddenhagen 1407101 – FSU	3,857,280	3,079,000	751,962
R. decurrens Chapman	Loran Anderson 23273 – FSU	6,931,756	5,589,372	688,873
R. distans (Michaux) Vahl	Chris Buddenhagen 14082120 – FSU	4,894,668	4,030,492	529,996
R. divaricata (Ham.) M.T.Strong	Robert Naczi 12107 - NY	4,478,092	3,776,406	858,680

R. elatior Kunth	Wayt Thomas 16400 – NY	5,307,938	4,415,256	877,377
R. emaciata (Nees) Boeckeler	Jeremy Bruhl 2320 – NE	5,657,984	4,649,328	556,786
R. exaltata Kunth	Wayt Thomas s.n – NY	5,151,874	4,088,012	416,693
R. eximia (Nees von Esenbeck) Boeckeler	Chris Buddenhagen 14082224 – FSU	3,591,702	3,072,912	811,811
R. fernaldii Gale	Richard Carter 21415 – VSC	5,685,332	4,628,466	646,765
R. gaudichaudii (Brongn.) L.B. Sm.	Wayt Thomas 16319 - NY	3,129,884	2,702,174	713,956
R. gigantea Link	Wayt Thomas 16407 – NY	4,339,182	3,794,984	384,450
R. glaziovii Boeckeler	M Reginato 1486 – NY	1,814,902	1,563,696	173,544
R. globosa (Kunth) Roem. & Schult.	Jeremy Bruhl 2319 - NE	3,713,422	3,158,265	199,121
R. globularis (Chapm.) Small var. *pinetorum* (Britton & Small) Gale	Chris Buddenhagen 1107084 - FSU	5,178,502	4,250,708	592,819
R. grayi Kunth	Chris Buddenhagen 1305106 – FSU	5,853,066	4,812,612	482,515
R. harperi Small	Chris Buddenhagen 14082233 – FSU	4,417,018	3,602,654	696,277
R. holoschoenoides (Rich.) Herter	Wayt Thomas 16306B – NY	5,493,150	4,698,590	795,612
R. imeriensis (Kük.) W.W.Thomas	Julian Aguirre 1899 – NY	5,636,852	4,779,506	971,455
R. inundata (Oakes) Fernald	Chris Buddenhagen 13101032 – FSU	2,813,228	2,382,076	443,639
R. latifolia (Baldwin ex Elliott) W.W.Thomas	Loran C. Anderson 27371 – FSU	4,095,978	3,578,968	881,513
R. longisetis R.Br.	K.L. Wilson 9684 – NSW	3,991,260	3,530,476	280,947
R. macra (C. B. Clarke ex Britton) Small	Wayt Thomas 14709 – NY	6,062,476	4,897,268	980,680
R. macrostachya Torr. ex A.Gray	Robert Naczi 12032 – NY	4,744,770	4,060,686	676,729
R. marisculus Lindl. & Nees	Jeremy Bruhl 2328 – NE	6,440,630	5,262,726	773,054
R. megalocarpa A.Gray	Chris Buddenhagen 1107085 – FSU	3,599,538	2,876,908	884,916
R. mixta Britton ex Small	Loran Anderson 25574 – FSU	5,348,182	4,419,824	515,468
R. nitens (Vahl) A.Gray	Loran C. Anderson 25687 – FSU	3,103,108	2,259,526	339,281
R. nivea Boeckeler	Taylor et al 2932 – BRIT	4,394,970	3,842,932	1,020,456
R. pallida M.A.Curtis	Wayt Thomas 14707 – NY	6,183,336	5,172,608	872,903
R. panicoides Schrad. ex Nees	ER Guaglione & Sobral 19990724 – NY	5,179,490	4,407,576	300,769
R. perplexa Britto n ex Small	Loran Anderson 25485 – FSU	5,590,824	4,454,782	1,153,277
R. pilulifera Bertol.	M Reginato 1476 – NY	3,595,134	3,013,238	462,259
	M Reginato 1482 – NY	4,207,248	3,674,720	1,329,031
R. polyphylla (Vahl) Vahl	Wayt Thomas 16483 – NY	4,324,978	3,770,886	1,251,776
R. pubera (Vahl) Boeckeler	ACG Costa sn – NY	4,164,982	3,653,106	2,797,919
R. punctata Elliott	Richard Carter 21712 – VSC	4,127,092	3,447,418	634,965
R. racemosa C.Wright ex Sauvalle	Wayt Thomas 14866 - NY	2,992,722	2,630,060	728,267
R. radicans (Schltdl. & Cham.) H.Pfeiff.	Wayt Thomas 14907 – NY	4,136,684	3,498,448	926,007
R. rariflora (Michaux) Elliott	Chris Buddenhagen 14082219 - FSU	4,362,706	3,714,820	873,893
R. recognita (Gale) Kral	Chris Buddenhagen 14082229 – FSU	4,382,002	3,620,784	582,753
R. riedeliana C.B.Clarke	Wayt Thomas 16401 - NY	5,251,760	4,372,724	1,258,905
R. riparia (Nees) Boec keler	Wayt Thomas 16153 – NY	4,535,984	3,925,288	391,118

R. robusta (Kunth) Boeckeler	Wayt Thomas 16403 – NY	3,577,870	3,230,414	650,633
R. rubra (Lour.) Makino	Jeremy Bruhl 2487 – NE	6,092,642	5,014,422	811,006
R. schiedeana (Schltdl.) Kunth	Wayt Thomas 16473 – NY	2,891,070	2,591,544	721,677
R. scirpoides (Torrey) A. Gray	Robert Naczi 12036 – NY	4,925,750	3,747,988	523,231
R. seslerioides Griseb.	Wayt Thomas 14924 – NY	4,916,342	3,991,846	813,824
Rhynchospora "sierrensis"	Wayt Thomas 16477 – NY	3,430,562	2,810,674	951,392
R. tenuis Link	Wayt Thomas sn – NY	4,826,084	3,829,980	661,128
R. tenuiflora (Brongn.) L.B. Sm.	Wayt Thomas 16317 – NY	4,342,424	3,722,124	261,844
R. wrightiana Boeckeler	M Reginato 1479 – NY	3,356,830	2,941,586	196,356
Scirpodendron ghaeri (Gaertn.) Merr.	Richard Carter 21416 - VSC	8,000,000	6,058,412	447,408
	Chrissie J. Prychid 47 – NE	-	-	-

Supplementary Table 2 – Name, monomer length, GC content and consensus sequence of shared satDNAs uncovered in the comparative clustering analysis, Tyba variants uncovered in the Tyba-like clustering analysis and of the Tyba consensus uncovered in the individual species clustering analysis.

Shared satDNAs			
Satellite	**Monomer lenght**	**GC content**	**Consensus**
CL01 (Tyba A)	172	27.3	CTAAGTCATTTCATCACAATAATCTACATTTAAACTCTTTATACTGTCTAGA ATATGATTCATATGTTATTTATTCAAAAAGATTGGATTATACATGGTAATT ACGCATATAAAGTGCAAATAATGCAATTCTGAGCATCATTTGAGCATTTCA ATCGTTTCTGGATTCATT
CL38 (Tyba B)	170	29.4	TGAAGTACTTAGATCACAAAACTTGATATTTATGCTCTACTTATGATATAA AAGAACTTTAAAGTCTTATATATTCATATTATTGAATTATACATCTATTTTG AGCATAAAGTGTACTTATAGCTCAATCTGAGCACCCGTTGTGCATTTTAAG CGTTTCCAGCCACATT
CL141 (Tyba B)	172	30.2	TGAAGTCCGTAGATCACAAAAATTAACATTTGAACTCCACATATGATACCA AATGACATTCAGATCTTACATATTCTAAATATTGAATTATATTTGTATTTAG AGCTTATGAAGTACATTTAAAGCTCATTTTGAGCACCTGTTGAGCATTTTA AGCGATTCTAGCCACATT
CL24 (Tyba C)	171	35.1	TTAAGTCATTTGATCGCAATAATCTATAGTTAAGCTCTTCATACAGTCTAG AATATGTTTATGAACTCACTTATTCAAAACGATTGGATTATACGCTGTAAT TACGCATAAACGCGCCGATTATGCAGATTCTGGGTACATTAGAGCATTTCG AGCGATTGTGCATTCATT
CL41 (Tyba C)	172	32.6	TTAAGTCATTTGATCGCAATAATCTATATTTAAGCACATTATACAGTGTAG AATAAGTTTATGATCTCAAAATTTCAAAAGGATTGGATTATACGCTGCAAT TAAGCACATAAAGTGCCAAAATGAGCGTTCTGGACAGAATTGGAGCATTT CAAGCGATTGTGCATTCATT
CL134 (Tyba C)	172	27.3	CTAAGTCATTTCATCACAATAATCTACATTTAAACTCTTTATACTGTCTAGA ATATGATTCATATGTTATTTATTCAAAAAGATTGGATTATACATGGTAATT ACGCATATAAAGTGCAAATAATGCAATTCTGAGCATCATTTGAGCATTTCA ATCGTTTCTGGATTCATT
CL168 (*Rg*Sat)	185	42.2	CATTGCTCAATTCTAATCGAGCGGTCCAAGAACGAACGAAGAATGCGGGA AGAGTAGTGTTCGAATTGGCTGAAATAGAGGTTCTGAGTATCTTTTAACTA ATTCAAGCGTCATTTCAACTCATTATCCCGCAAGAAATGACTCATTAGAAT CGTCTTAGCACCGACAACATCTTCCTAGCTCGG
CL332 (*Rg*Sat)	184	35.3	CTTAACGAAGAGAGAATTAGTGCTAAAAGAGTGTTGTTTCTGCTGAAATA GTGATTTTGAGTATGTTTCATCGATTTCAAGTGTTATTTCAACTCATTAGCT AGCAAGAAAGTACTCATTAGAATCCTCGTGATACGTACATCATCATCTATG CTCGATTCGTCTACATTCTAATCGAATGGTG
CL06	182	31.3	TGGTATCTAGATTCTAAATATCAATTCTTGCATTCAAATGAGTTGAAATGA TGTTAAAATTGCACAAAACAAACTTAGAAGCTGTATTTCAGCTTAAATTGC ATTGTTCTTTGCATTCTTGGTTGAATATTAACACGTACGATTACGTGGTAGA AAAATAGGACCCTAGAGGTTATTGATGA
CL10	446	62.8	ATGGGGTCTATTTCCCCACCCGGAGGGTGCGGGCTGCTAGGGGGGAGCAG CCGCCGGACCCCGGCACCCTCCTGAAAATCACTACCCCCTAATAGGGAAA ACTGATTTTTTTGTAACTTACCCCTAGGGGGTTACCATCCCCTGGATTTCCC CCATGCACAAAAACGCTGTGCCGCTGGACCAGCGAGCCGGCTCCAATGGT CCAGCGGAGAGTCCAATGGTCCGGCGGCCGGCACCGTGGTTCCAGGAGCC GGCACCCATTTTCCGTCCTGGCCGGCACCCCTGGGCACGATTTTCCCTAGC CCAGTGCCTTTTGCACTGGGTACATTTTCCCTCGCACAGGGTCATTTTTTCC GACGCCCAATCCGACAAGGCCTCCCCAGGGTCCAGCCGCCGGCTCCGGAG CGGAGGGGGGGAGGACCCTCGTTCCATCCTAGCCGGCCTCA
CL61	59	32.2	AAAACTCAGTTTTTTCAAAAACACTAAAAAAATGAGTTTACGTTCGGCCGA ACGTGAAA
CL165	169	55.0	TTCAAAAACGTTCCGGATCCCTAGCCGGCGGGTCACAGGCCCGGAGGTCC CGAAACCTCCGGGCGCCGACCTGGCCCGGGTGAGTCCAGTACATTCATTTC CATCCGAATTAATGCTGGGAGTACGAAAAAGCGTGAAAAAAGATCCACAG GTGCAAAAACAGTTGCAG
CL239	347	51.3	CCCTCGAGAGATCCCCCGTCGGCCGAAGCGGCCCTTGGCACGAGCGCACC ATCGTGCACACCCCCGCCAGCCGGACCAGGAAATGTAATATGAGCTTTCC AGTGTCATTTCGGAGCACCAGAAGACGTTTTTTACCCCATAAAGCGACGA

CL261	189	43.9	ATTTTCCAAAAATTCAAAAATTTTGTCGTAGAGCTCTGGGTTAGCTTCAATTTGATAGTTCACTTGCGAGATTTGGAGTTTTGTGCCAAAAGTTACGCCCGATTTAAGCACGAAAGGTTAGTTTCTCATTTCGTGTCAAAAGCTATGCTCCCTCGAGCGCTCCTCCTCTCGCGAGAGCGTTCCCTAGCACGGCCGGC
CL319	184	45.7	TTTTTGAAAAAAGAGGCAAAAAACCGCCACATCTATTGAAAAGAAGTAAAACTCCCAATACATCAGGTTTCGGACTTTGGGTCGCGAAGCTCCTGTCAAAATTTCGGCCCGATCCGACGGTCGGATCGAAAGTTCGGCCCCATTTTGTGAAGTGTGCGCGGGACCACATTGCATCAAATTTTTTATTTT
CL370	59	50.8	ATAAGTACATATCCGAAAAAGGAAAATGCTCAATAGGCGGAGCGCTCCGAGAGCGCTCGGCACGGTCGAATTGCTCGTTTCTCGAGCTTTCCGTATTGGTGTCGTACGTCTCGATCGTAGCTACGAGCAGAAAGTTATGATCGTTTTAGGATTGGAAACTCGAATACATGCAACATATACATAC
CL376	59	52.5	TCACCCGGCCCAAAAATCTCAATTTTTCGTGACCCGCCGGCTATCAAAATTCGTCCGAG
			CAAAATTCTCATTTTTTCGTGACCGGCCGGCTAGGAGAATCTTTCCGAGTCCGCCGGTG

***Tyba* variants obtained in the *Tyba*-like clustering analysis**

Satellite	Monomer lenght	GC content	Consensus
TL-CL01 (*Tyba* A1)	172	31.4	TGAAGTCCTTTAATCGCAAAAATTGATATTTAAGCTCCTCATATGATATAAAACAACATTAAGAACTTACTTATTCGAATTAATGAATGATATATCTGTTTTTTGCGTAAAATGTACCGTTATTGCTCATTTACAGCACCTGTAGTGCATTTTAAGCGATTCTGGCCTCGTT
TL-CL04 (*Tyba* A2)	172	33.1	TGAAGTCCTTTAAGCGCAATAATTGATGTTTAAACTCCCCATATGATATAAAACATAATTAAGAACTTAGTTACTCGAATTCATGAATGATATATATGTTTTTTGCGCAAAATGTACCGTTTGGGCTGATTTAAAGCCGTTGTAGTGCATTTTAAGCGATTCTGGACTCGTT
TL-CL05 (*Tyba* A3)	172	34.9	TGAGGTCATTTAATCGCACAAATCGATATTTAAGCTCCTCATATGATATAAAACAACATTAAGGACTTATTTATCCGAATTAATGAGTGATATATCTTCTTTTTGCGCAAAATGTACCGTTATTGCTCAATTACAGCACCTGTAGTGCGTTTTACGCGATTCTGGACTCGTT
TL-CL12 (*Tyba* A4)	173	36.4	CGGAGTCCTTTAATCGCACAAATTGACATTTAAGCTCCTCATATGGTATAAAACAATGTTAAGAACTTACATATCCGAATTAATGAGCCGTATATCTGTTATTTGCGCAAAATGCACCGTTATTGCTCCTTTACAGCACATTTTAGTGCATTTTAATGCATTCTGGCCTCGTT
TL-CL02 (*Tyba* B1)	172	29.1	TGAAGTCCGTAGATCACAAAAATTAACATTTGAACTCCACATATGATACCAAATAACATTCAGATCTTACATATTCTAATTATTGAATTATATTTGTATTTAGAGCTTATAAAGTACATTTAAAGCTCATTTTGAGCACCTGTTGAGCATTTTAAGCGATTCTAGCCACATT
TL-CL08 (*Tyba* B2)	170	33.5	TGAAGTTCTTTGATCGCAAAACTTGATATTTATGCTCTACTTATGATACTAAAGGACTTTAAAGTCTTACATATTCATATTATTGAATTATACATCTATTTTGAGCATATAGTGTACTTATAGCTCAATCTGAGCACCCGTTGTGCATTGTGAGCGTTTCCCGCCACAAT
TL-CL09 (*Tyba* B3)	172	30.2	TGAAGTCCTTAGATCACAAAAATCTATATTTAAGCTCCACATATGATCTACAATAACATTATGAACTCAATTCTTCAAATTATTAATTGAGACACATATTTTGTGCACATAATTGCATTCAAATAGCTTTTTTCTGCACCTGATGAGGTTTTTAAGCGTTTCTAGCCACATT
TL-CL03 (*Tyba* C1)	171	35.1	TTAAGTCATTTGATCGCAATAATCTATAGTTAAGCTCTTCATACAGTCTAGAATATGTTTATGAACTCACTTATTCAAAACGATTGGATTATACGCTGTAATTACGCATAAACGCGCCGATTATGCAGATTCTGGGTACATTAGAGCATTTCGAGCGATTGTGCATTCATT
TL-CL06 (*Tyba* C2)	173	31.2	TTAAGTCATTTGATCACAATAATCTATATTTAAGCTCTTTATACAGTGGAGAATAAGTTTATGATCTTAAAAATTCAAAAGGATTGGATTATACATGGCAATTACGCACATAAAGTGCAAATAATGAGCGTTCTGGACAGAATTGGAGCATTTCAAGCGATTGTGCATTCATT
TL-CL07 (*Tyba* C3)	172	26.7	CTAAGTCATTTCATCACAATAATCTACATTTAAACTCTTTATACTGTCTAGAATATGATTCATATGTTATTTATTCAAAAAGATTGGATTATACATGGTAATTACGCATATAAAGTGCAAATAATGCAATTCTGAGTATCATTTGAGCATTTCAATCGTTTCTGGATTCATT

TL-CL10 (*Tyba* C4)	173	25.4	TTAAGTCATTTGATCACAATAATCTATATTTAAGCTCTTCATATAGTCTAGA ATATGTTTATGATCTTACTAATTCAAAATGATTGGATTATACAAAGTAATT ACGCATATAATGTTCAAGTAATGCAAATTCTGAGTATCATTTGAACATTTC AAGCGATTTTGCATTCATT
TL-CL11 (*Tyba* C5)	174	35.1	TTAAGTGATTTGATCGCAATAATCTATATTTAAGCTCTTGATAGAGTGGAG AATAAGATTATGATCTTAAAATTTCAAGAGGATGTGGTTATACTCGTGAAT TACGCGCATAAAGTGCAACAAATTCACGTTCTGGGCAGAATTGGAGTGTTT TGAAGCCATTGTGCATTCGTT

Consensus sequence of the most abundant *Tyba* for each species

Species	Monomer lenght	GC content	Consensus
R. alba	172	30.8	TTGAAGTCCTTTAATCGCAAAAATTGATATTTAAGCTCCTCATATGATATA AAACAACATTAAGAACTTAATTATTCGAATTAATGAATGATATATCTGTTT TTTGCGTAAAATGTACCGTTATTGCTCATTTACAGCACCTGTAGTGCATTTT AAGCGATTCTGGCCTCGT
R. barrosiana	172	30.8	TTGAAGTCCTTTAATCGCAAAAATTGATATTTAAGCTCCTCATATGATATA AAACAAAATTAAGAACTTACTTATTCGAATTAATGAATGATATATCTGTTT TTTGCGTAAAATGTACCGTTATTGCTGATTTACAGCACCTGTAGTGCATTTT AAGCGATTCTGGCCTCGT
R. berteroi	173	25.4	CTAAGTCATTTCATCACAATAATCTACATTTAAACTCTTTATACTGTCTAGA ATATGATTCATATCTTATTTATTCAAAAAGATTGGATTATACATGGTAATT ACGCATATAAAGTGCAAATAATGCAATTCTGAGTATCATTTGAGCATTTCA ATCGTTTCTGGATTCATT
R. brachychaeta	172	33.7	TTGAGGTCATTTAATCGCACAAATCGATATTTAAGCTCCTCATATGATATA AAACAACATTAAGAACTTATTTATACGAATTAATGAGTGATATATCTTCTT TTTGCGCAAAATGTACCGTTATTGCTCAATTACAGCACCTGTAGTGCGTTT TACGCGATTCTGGACTCGT
R. brasiliensis	172	31.4	TTGAAGTCCTTTAATCGCAAAAATTGATATTTAAGCTCCTCATATGATATA AAACAAAATTAAGAACTTACTTATTCGAATTAATGAATGATATATCTGTTT TTTGCGTAAAATGTACCGTTATTGCTGATTTACAGCACCTGTAGTGCATTTT AAGCGATTCTGGCCTCGT
R. bucherorum	172	31.4	TTGAAGTCCTTTAATCGCAAAAATTGATATTTAAGCTCCTCATATGATATA AAACAACATTAAGAACTTACTTATTCGAATTAATGAATGATATATCTGTTT TTTGCGTAAAATGTACCGTTATTGCTCATTTACAGCACCTGTAGTGCATTTT AAGCGATTCTGGCCTCGT
R. californica	172	31.4	TTGAAGTCCTTTAATCGCAAAAATTGATATTTAAGCTCCTCATATGATATA AAACAACATTAAGAACTTACTTATTCGAATTAATGAATGATATATCTGTTT TTTGCGTAAAATGTACCGTTATTGCTGATTTACAGCACCTGTAGTGCATTTT AAGCGATTCTGGCCTCGT
R. capitellata	172	31.4	TTGAAGTCCTTTAATCGCAAAAATTGATATTTAAGCTCCTCATATGATATA AAACAACATTAAGAACTTACTTATTCGAATTAATGAATGATATATCTGTTT TTTGCGTAAAATGTACCGTTATTGCTCATTTACAGCACCTGTAGTGCATTTT AAGCGATTCTGGCCTCGT
R. cephalotes	172	29.6	TTGAAGTCCTTAGATCACAAAAATCTATATTTAAGCTCCACATATGATCTA CAATAATATTATGAACTCAATTCTTCAAATTATTAATTGAGACACATATTTT GTGCACATAATTGCATTCAAATAGCTTTTTTCTGCACCTGATGAGGTTTTTA AGCGTTTCTAGCCACAT
R. colorata	172	26.7	CTAAGTCATTTCATCACAATAATCTACATTTAAACTCTTTATACTGTCTAGA ATATGATTCATATCTTATTTATTCAAAAAGATTGGATTATACATGGTAATT ACGCATATAAAGTGCAAATAATGCAATTCTGAGTATCATTTGAGCATTTCA ATCGTTTCTGGATTCATT
R. crinipes	172	31.4	TTGAAGTCCTTTAATCGCAAAAATTGATATTTAAGCTCCTCATATGATATA AAACAACATTAAGAACTTACTTATTCGAATTAATGAATGATATATCTGTTT TTTGCGTAAAATGTACCGTTATTGCTCATTTACAGCACCTGTAGTGCATTTT AAGCGATTCTGGCCTCGT
R. decurrens	172	31.4	TTGAAGTCCTTTAATCGCAAAAATTGATATTTAAGCTCCTCATATGATATA AAACAACATTAAGAACTTACTTATTCGAATTAATGAATGATATATCTGTTT TTTGCGTAAAATGTACCGTTATTGCTCATTTACAGCACCTGTAGTGCATTTT AAGCGATTCTGGCCTCGT

R. distans	172	34.9	TTGAGGTCATTTAATCGCACAAATCGATATTTAAGCTCCTCATATGATATA AAACAACATTAAGGACTTATTTATCCGAATTAATGAGTGATATATCTTCTT TTTGCGCAAAATGTACCGTTATTGCTCAATTACAGCACCTGTAGTGCGTTT TACGCGATTCTGGACTCGT
R. emaciata	174	35.0	TTAAGTGATTTGATCGCAATAATCTATATTTAAGCTCTTGATAGAGTGGAG AATAAGATTATGATCTTAAAATTTCAAGAGGATGTGGTTATACTCGTGAAT TACGCGCATAAAGTGCAACAAATTCACGTTCTGGGCAGAATTGGAGTGTTT TGAAGCCATTGTGCATTCGTT
R. exaltata	170	29.4	TGAAGTACTTAGATCACAAAACTTGATAATTATGCTCTACTTATGATATAA AAGAACTTTAAAGTCTTATATATTCATATTATTGAATTATTTATCTATTTTG AGCATAAAGTGTACTTATAGCTCAATACGAGCACCTGTTGCGCGTTTTAAG CGTTTCCAGCCACATT
R. eximia	171	35.0	TTAAGTCATTTGATCGCAATAATCTATAGTTAAGCTCTTCATACAGTCTAG AATATGTTTATGAACTCACTTATTCAAAACGATTGGATTATACGCTGTAAT TACGCATAAACGCGCCGATTATGCAGATTCTGGGTACATTAGAGCATTTCG AGCGATTGTGCATTCATT
R. gaudichaudii	172	28.5	TTGAAGTCCGTAGATCACAAAAATTAACATTTGAACTCCACATATGATACC AAATAACATTCAGATCTTACATATTCTAATTATTGAATTATATTTGTATTTA GAGCTTATAAAGTATATTTAAAGCTCATTTTGAGCACCTGTTGAGCATTTT AAGCGATTCTAGCCACAT
R. glaziovii	170	30.0	TGAAGTACTTAGATCACAAAACTTGACATTTATGCTCTACTTATGATATAA AAGAACTTTAAAGTCTTATATATTCATATTATTGAATTATTTATCTATTTTG AGCATAAAGTGTACTTATAGCTCAATACGAGCACCTGTTGCGCGTTTTAAG CGTTTCCAGCCACATT
R. globularis	172	30.3	TTGAAGTCCTTTAAGCGCAATAATTGATGTTTAAACTCCCCATATGATATA AAACATAATTAAGAACTTAGTTACTCGAATTCATGAAGGATATATATGTTT TTTGCGCAAAATGTACCGTTTGGGCTGATTTAAAGCCGTTGTAGTGCATTT TAAGCGATTCTGGACTCGT
R. grayi	172	29.7	TTGAAGTCCTTTAATCGCAAAAATTGATATTTAAGCTCCTCATATGGTATA AAACAACATTAAGAACTTAATTATTCGAATTAATGACTGATATATCTGTTT TTTGCGTAAAATGTACCGTTATTGCTCATTTACAGCACAATTAGTACATTTT AAGCGTTTCTGGACTCGT
R. harperi	172	33.7	TTGAGGTCATTTAATCGCACAAATCGATATTTAAGCTCCTCATATGATATA AAACAACATTAAGAACTTATTTATACGAATTAATGAGTGATATATCTTCTT TTTGCGCAAAATGTACCGTTATTGCTCAATTACAGCACCTGTAGTGCGTTT TACGCGATTCTGGACTCGT
R. imeriensis	172	30.2	TTGAAGTCCTTAGATCACAAAAATCTATATTTAAGCTCCACATATGATCTA CAATAACATTATGAACTCAATTCTTCAAATTATTAATTGAGACACATATTT TGTGCACATAATTGCATTCAAATAGCTTTTTTCTGCACCTGATGAGGTTTTT AAGCGTTTCTAGCCACAT
R. latifolia	173	31.2	TTAAGTCATTTGATCACAATAATCTATATTTAAGCTCTTTATACAGTGGAG AATAAGTTTATGATCTTAAAAATTCAAAAGGATTGGATTATACATGGCAAT TACGCACATAAAGTGCAAATAATGAGCGTTCTGGACAGAATTGGAGCATT TCAAGCGATTGTGCATTCATT
R. macra	172	30.8	TTGAAGTCCTTTAATCGCAAAAATTGATATTTAAGCTCCTCATATGATATA AAACAACATTAAGAACTTAATTATTCGAATTAATGAATGATATATCTGTTT TTTGCGTAAAATGTACCGTTATTGCTCATTTACAGCACCTGTAGTGCATTTT AAGCGATTCTGGCCTCGT
R. marisculus	172	31.4	TTGAAGTCCTTTAATCGCAAAAATTGATATTTAAGCTCCTCATATGATATA AAACAACATTAAGAACTTACTTATTCGAATTAATGAATGATATATCTGTTT TTTGCGTAAAATGTACCGTTATTGCTGATTTACAGCACCTGTAGTGCATTTT AAGCGATTCTGGCCTCGT
R. megalocarpa	172	35.4	TTGAAGTCCTTTAATCGCAAAAACTGACATTTAAGCTCCTCATATGATATA AAACAAAGTTAAGAGCTTACTTATTCGAATTAATGAATGATATATCTGTTT TTTGCGCAAAATGCGCCGTTATTGCTGATTTAGAGCACCTGTTGTGCGTTTT AAGCGATTCGAGCCTCGT
R. mixta	172	31.4	TTGAAGTCCTTTAATCGCAAAAATTGATATTTAAGCTCCTCATATGATATA AAACAACATTAAGAACTTACTTATTCGAATTAATGAATGATATATCTGTTT TTTGCGTAAAATGTACCGTTATTGCTGATTTACAGCACCTGTAGTGCATTTT AAGCGATTCTGGCCTCGT

R. nitens	171	35	TTAAGTCATTTGATCGCAATAATCTATAGTTAAGCTCTTCATACAGTCTAG AATATGTTTATGAACTCACTTATTCAAAACGATTGGATTATACGCTGTAAT TACGCATAAACGCGCCGATTATGCAGATTCTGGGTACATTAGAGCATTTCG AGCGATTGTGCATTCATT
R. nivea	172	26.7	CTAAGTCATTTCATCACAATAATCTACATTTAAACTCTTTATACTGTCTAGA ATATGATTCATATCTTATTTATTCAAAAAGATTGGATTATACATGGTAATT ACGCATATAAAGTGCAAATAATGCAATTCTGAGTATCATTTGAGCATTTCA ATCGTTTCTGGATTCATT
R. pallida	172	30.8	TTGAAGTCCTTTAATCGCAAAAATTGATATTTAAGCTCCTCATATGATATA AAACAACATTAAGAACTTAATTATTCGAATTAATGAATGATATATCTGTTT TTTGCGTAAAATGTACCGTTATTGCTCATTTACAGCACCTGTAGTGCATTTT AAGCGATTCTGGCCTCGT
R. panicoides	172	30.2	TTGAAGTCCGTAGATCACAAAAATTAACATTTGAACTCCTCATATGATACC AAATGACATTCCAATCTTATATATTCTAAATATTGAATAATATTTGTATTTA GAGCTCATGAAGTACATTTAATGCTCATTTTGAGCACCTGTTGAGCATTTT AAGCGATTCTAGCCACAT
R. perplexa	172	30.8	TTGAAGTCCTTTAATCGCAAAAATTGATATTTAAGCTCCTCATATGATATA AAACAAAATTAAGAACTTACTTATTCGAATTAATGAATGATATATCTGTTT TTTGCGTAAAATGTACCGTTATTGCTCATTTACAGCACCTGTAGTGCATTTT AAGCGATTCTGGCCTCGT
R. pilulifera	172	28.5	TTGAAGTCCGTAGATCACAAAAATTAACATTTGAACTCCACATATGATACC AAATAACATTCAGATCTTACATATTCTAATTATTGAATTATATTTGTATTTA GAGCTTATAAAGTATATTTAAAGCTCATTTTGAGCACCTGTTGAGCATTTT AAGCGATTCTAGCCACAT
R. pilulifera 2	172	28.5	TGAAGTCCGTAGATCACAAAAATTAACATTTGAACTCCACATATGATACCA AATAACATTCAGATCTTACATATTCTAATTATTGAATTATATTTGTATTTAG AGCTTATAAAGTATATTTAAAGCTCATTTTGAGCACCTGTTGAGCATTTTA AGCGATTCTAGCCACATT
R. polyphylla	171	28.7	TTGAAGCACTTAAATCAAAAAAATCTATATTTAAGCTCCACTTATAATCTT AGATAACGTTATGAAGTCATTTATTCAAATTATTAATTGAGACACACATTT TGTGCAAATAAGTACAATCAACTAGCTGTTTCTGCAGCTAATGATGTTTTC AAGCATTTCTAGCCACAT
R. pubera	172	26.2	TTAAGTCATTTGATCACAATAATCTATATTTAAGCTCTTTATATAGTCTAGA ATATGTTTATGAACTCACTTATTCAAAACGATTGGATTATACATTGTAATT ATTCACATAATGTGCAGATTTTACAGATTCTGAGTATATTTGAGCATTTCA AGCGATTTTGCATTCATT
R. punctata	173	34.1	TTGAAGTCCTTTAAGCGCAATAATTGATGTTTAAACTCCCCATATGGTCTA AAACATAATTAAGAACTTAGTTACTCGAATTCATGAAGGATATATATGTTT TTTGCGCAAAATGTACCGTTTTGGGCTGATTTAAAGCCGTTCTAATGCATTT TAAGCGATTCTGGACTCGT
R. radicans	172	27.9	CTAAGTCATTTCATCACAATAATCTACATTTAAACTCTTTATTCTGTCTAGA ATACGATTCATATGTTATTTATTCAAAAAGATTGGATTATACATGCTAATT ACGCATATAAAGTGCAAATAATGCAATTCTGAGCATCATTTGACCATTTCA ATCGTTTCTGGATTCATT
R. rariflora	172	31.4	TTGAAGTCCTTTAATCGCAAAAATTGATATTTAAGCTCCTCATATGATATA AAACAACATTAAGAACTTACTTATTCGAATTAATGAATGATATATCTGTTT TTTGCGTAAAATGTACCGTTATTGCTCATTTACAGCACCTGTAGTGCATTTT AAGCGATTCTGGCCTCGT
R. recognita	173	34.1	TTGAAGTCCTTTAAGCGCAATAATTGATGTTTAAACTCCCCATATGGTCTA AAACATAATTAAGAACTTAGTTACTCGAATTCATGAAGGATATATATGTTT TTTGCGCAAAATGTACCGTTTTGGGCTGATTTAAAGCCGTTCTAATGCATTT TAAGCGATTCTGGACTCGT
R. riedeliana	171	25.8	TGAATTCATTTTAACATAAAAATTGATATTTAGATACTACATATGATCTCA ATAATATTCAAGAATTATTCTTGCGAATTATTAGATCAGATATTTATTTGG AGCATCTGAAATGTATTTACAGCTTATTTTGTGCACCTGATGTACATTTCAA GCAATTCTTGCTACATT
R. riparia	172	29.6	TTAAGTCATTTGATCGCAATAATCTATATTTAAGCACATTTTACAGTTTATA ACAAGTTTATGATCTTAAAATTTGAAAAGGATTGGATTATACACTGCAAAT AAGCATATAAAGTGCAAAAATGAGAGTTCTGGACAGAATTGGAGCATTTC AAGCGATTGTGCATTCGTT

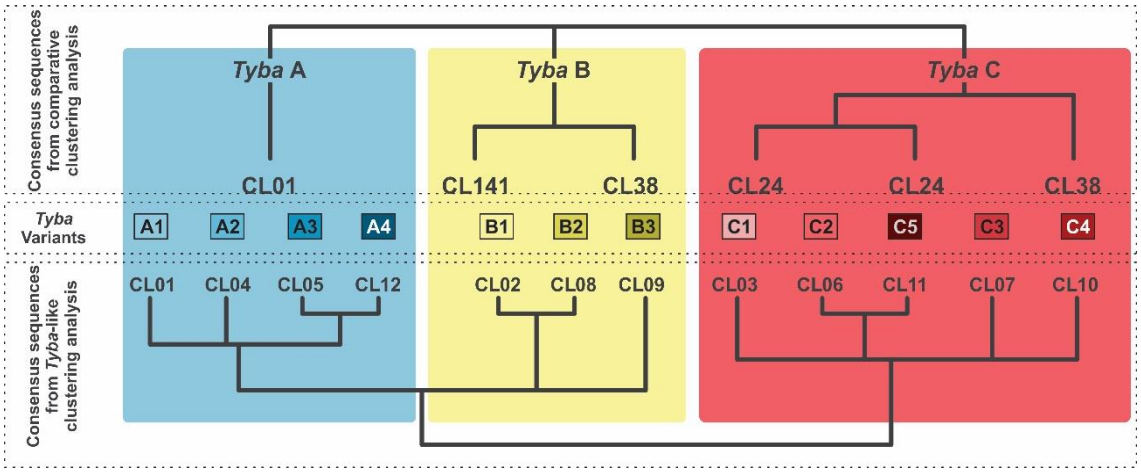
R. robusta	171	35	TTAAGTCATTTGATCGCAATAATCTATAGTTAAGCTCTTCATACAGTCTAG AATATGTTTATGAACTCACTTATTCAAAACGATTGGATTATACGCTGTAAT TACGCATAAACGCGCCGATTATGCAGATTCTGGGTACATTAGAGCATTTCG AGCGATTGTGCATTCATT
R. schiedeana	171	29.9	TTGAAATTCTTAGATCACAAAAACCTTCATTTAAGCTCCACTTATAATGTA AGATAACGTTATGAAGTCATTTATTCAAATTATTAATTGAGACACACATTT TGTGCAAATAAGTACAATCAAGTAGCTGTTTCTGCAGCTAATGATGTTTTC AAGCATTTCTAGCCACAT
R. scirpoides	171	35	TTAAGTCATTTGATCGCAATAATCTATAGTTAAGCTCTTCATACAGTCTAG AATATGTTTATGAACTCACTTATTCAAAACGATTGGATTATACGCTGTAAT TACGCATAAACGCGCCGATTATGCAGATTCTGGGTACATTAGAGCATTTCG AGCGATTGTGCATTCATT
R. seslerioides	173	30.6	TTAAGTCATTTGATCACAATAATCTATATTTAAGCTCTTTATACAGTGGAG AATAAGTTTATGATCTTAAAAATTCAAAAGGATTGGATTATACATGGCAAT TACGCACATAAAGTGCAAATAATGAGCGTTCTGGACAGAATTGGAGTATT TCAAGCGATTGTGCATTCATT
R. tenuiflora	172	28.5	TTGAAGTCCGTAGATCACAAAAATTAACATTTGAACTCCACATATGATACC AAATAACATTCAGATCTTACATATTCTAATTATTGAATTATATTTGTATTTA GAGCTTATAAAGTATATTTAAAGCTCATTTTGAGCACCTGTTGAGCATTTT AAGCGATTCTAGCCACAT
R. tenuis	173	34.1	TTAAGTCATTTGACCGCAATAATCTAGATTTAAGCTCTTTATACAGTGGAG AATAAGTTTATGATCTCACTTTTTCAAGAGGATTGGACTATAGGGGGTAAT TACGCACATAAAGTGCAAAAAACGCACGTTCTGAGCAGAATTTTGAGCAT TTCAAGCGATTGTGGGTTCGTT
R. wrightiana	172	34.9	TTGAGGTCATTTAATCGCACAAATCGATATTTAAGCTCCTCATATGATATA AAACAACATTAAGGACTTATTTATCCGAATTAATGAGTGATATATCTTCTT TTTGCGCAAAATGTACCGTTATTGCTCAATTACAGCACCTGTAGTGCGTTT TACGCGATTCTGGACTCGT

Supplementary Table 3 – Proportion of different repetitive DNA classes per species identified in individual clustering analyses.

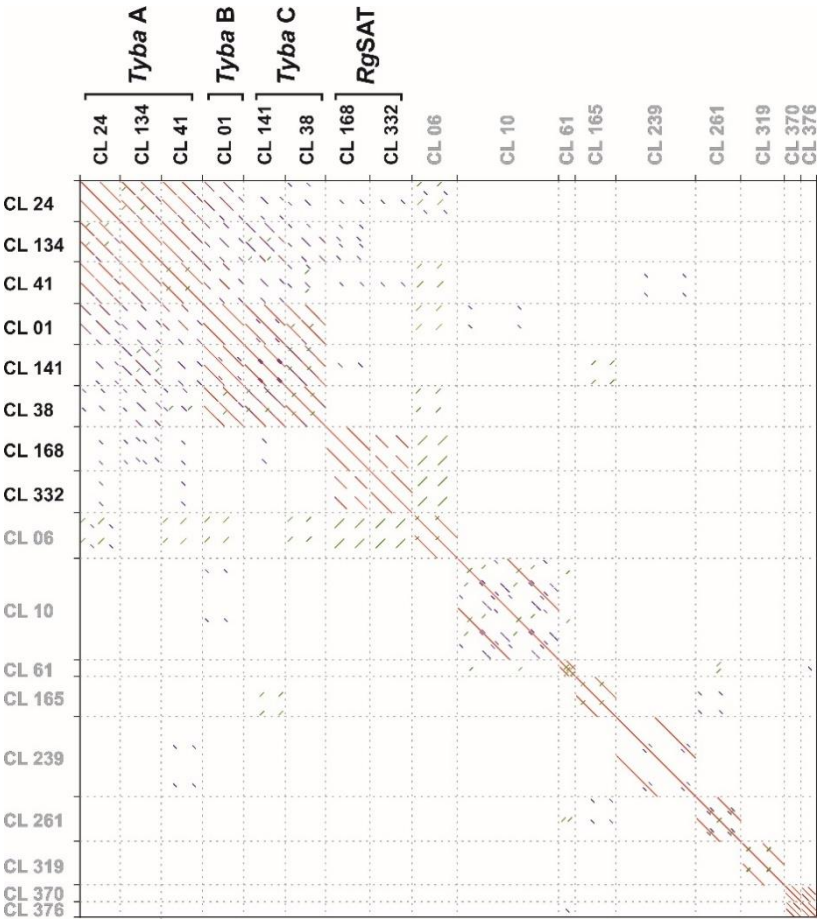
Species	Satellite DNA	rDNA 5S	rDNA 35S	Unclassified LTR	LTR-Ty1 copia	LTR-Ty3 gypsy	Pararetrovirus	LINE	Class II/ Subclass I	Class II/ Subclass II	Unclassified
R. affinis	0.13715124	0.05202516	1.68300393	2.75647611	4.01681685	7.38434129	0	0.17005435	1.1591917	0	7.281148167
R. alba FSU	3.15263458	0.06612889	2.76255445	0	0.47401295	1.12552977	0	0.04819109	0.13975417	0	26.33014112
R. alba NE	2.11338517	0.1107781	4.20215298	0.07058953	0.90149943	0.64094104	0	0.010974	0.13050165	0	26.2596004
R. albiceps	1.12454666	0.1383317	1.27472479	0.85960628	3.02725542	17.6226691	0.29344584	0.08366538	4.573831	0	12.59164013
R. barbata	1.86286728	0.0383454	7.9950151	0.2926233	5.56679289	20.845516	0.11887073	0	1.3650961	0	13.82471361
R. barrosiana	2.88889385	0.05904807	1.95554074	0	0.43196941	1.32031477	0	0.01102231	0.36098051	0.010235	25.1207533
R. barrosiana 2	3.04155978	0.02823065	1.99223178	0	0.4003426	0.67561809	0	0.01161185	0.32342739	0.01736451	17.88257327
R. berteroi	1.80814518	0.04177907	0.71198079	0.92496213	4.03612409	0.95775206	0.22626075	0.0903	1.97556792	0	17.1643552
R. biflora	0.62159668	0.06701363	0.95865712	2.72691131	3.54556429	5.15298577	0.09427187	0	2.00823543	0.03522743	19.48276346
R. brachychaeta	2.04218394	0.01104613	0.7969169	0	0.44098605	1.30467067	0.03473394	0.03927513	0.1729333	0	22.47089652
R. brasiliensis	3.31570197	0.05323081	1.24162457	0.3315702	0.47009856	1.55427538	0	0.0266154	0.44155915	0	28.1302359
R. brownii	3.07041691	0.16649827	1.39865809	0.07151198	0.36191594	0.84422852	0	0.01222117	0.35235681	0.01137416	22.78497401
R. brownii 2	3.24408732	0.02628583	1.70623638	0.05101012	0.4018348	1.20589479	0	0.02732685	0.24333908	0	23.76928332
R. brownii 3	3.24248395	0.02627283	1.70539308	0.05098491	0.4016362	1.20529879	0	0.02731334	0.24321881	0	23.8069597
R. bucherorum	2.27683009	0.02368108	0.71246482	0.08509285	0.41813017	0.11381058	0	0.03843758	0.19112755	0.01908625	15.51596927
R. californica	3.24486766	0.02397051	2.95384308	0	0.70077558	1.07014673	0	0.0809206	1.15219346	0	28.98099573
R. capitellata	2.06082408	0.0533677	2.6364149	0.02882359	0.28433406	0.67238272	0	0.11743412	0.05575918	0	23.76259931
R. careyana	2.49195485	0.070677	0.87454726	0.070677	1.31987678	1.60322923	0.15209518	0	0.40687608	0	37.95762817
R. cephalotes	2.52099828	0.12635979	0.30143868	0.36909453	0.48977329	0.47049224	0.12722055	0.01463295	0.26459808	0	30.28847561
R. colorata	1.7991117	0.05621142	2.31759567	0.30275678	6.09899674	1.31721708	0.51305906	0.26062707	1.23053377	0	24.79616146
R. consanguinea	1.65903172	0.25704817	1.09214871	0.54607435	3.10905881	20.5092308	0.22950729	0.12515798	4.6098835	0	10.27917267
R. corniculata	2.6320298	0.07448319	1.22439913	0.26366759	2.43587622	2.327999	0.32755181	0	0.5338688	0	22.83265529
R. corymbosa 1	9.14270004	0.13927012	2.96839154	0.02028206	1.05556833	2.54787692	0.40699325	0	0.51516422	0	19.32744703
R. corymbosa 2	5.56427511	0.08012067	5.49566737	0.06343866	1.62214996	0.67033514	0.08411496	0	0.71427229	0	17.21889626
R. corymbosa 3	8.71429446	0.69154389	10.4882881	0	1.31882927	0.57373663	0.07611573	0	0.54543229	0	15.41668579
R. crinipes	3.1637237	0.16835957	1.61736896	0	0.18977023	0.92863735	0	0.01236765	0.15306625	0	26.62514861

R. decurrens	3.00635966	0.04383972	1.83720366	0.03890412	0.04848499	0.18493975	0	0	0.03454918	0	24.55546959
R. distans	2.8462479	0.03490592	2.25209247	0	1.10698949	1.10378192	0	0.01320765	0.33453083	0	25.09320825
R. divaricata	0.019332	0.01688638	1.17878604	1.24726324	4.68870825	2.98062142	0.07546467	0.34308473	0.3246844	0	20.88938836
R. elatior	0.23490472	0.19911623	1.58871272	2.42039625	1.60934239	4.01822706	0.22875001	0.11477392	1.81415743	0.02724029	24.91141208
R. emaciata	3.94316667	0.02191147	1.91815168	0.83802394	1.27804938	3.18560452	0.03502243	0.54778676	0.13326485	0	28.97702169
R. exaltata	3.69312659	0.16342967	1.30527751	0.66595791	2.0079531	2.13370515	0.01559901	0.10967307	0.64843902	0	33.52828101
R. eximia	2.67968776	0.04570029	3.52729884	0.9169622	2.06759948	2.50415429	0.10334918	0.05456935	0.10618235	0	21.93281441
R. fernaldii	0.2467666	0.03741699	1.09668891	0.29670746	2.09627917	2.82931204	0	0.01824465	0.65201426	0.01515234	20.21058653
R. gaudichaudii	3.7807652	0.06344929	0.50521321	1.35932747	3.01209038	1.61312462	0.43069881	0.13866401	1.08326003	0	14.80609449
R. gigantea	7.4269736	0.08141501	5.60202887	0.82949668	2.42944466	2.36207569	0.67238913	0.12459358	0.94368578	0	14.98634413
R. glaziovii	4.95839672	0.03687826	1.3022634	0.41660904	1.29938229	2.22421979	0.02996358	0.01325312	0.7652238	0	18.1141382
R. globosa	1.8621843	0.10295248	0.79248296	1.07623003	3.26836446	15.4956032	0.19586081	0.12806284	1.60003214	0	25.19573481
R. globularis var. pinetorum	2.82345876	0.04925618	3.45754775	0.05566623	0.23160526	0.75098808	0	0.10610321	0.03947242	0.2941876	25.7562595
R. grayi	5.4796224	0.04331472	2.2938147	0	0.25139115	1.17778722	0	0.01015512	0.07253661	0.79147799	22.60427137
R. harperi	3.46686735	0.02585178	1.9900126	0.12150337	1.32906875	1.42687465	0	0	0.56701571	0	21.7502517
R. holoschenoides	1.44203456	0.18777997	3.2354967	0.16653846	1.16702614	0.38712337	0.8436273	0.01860203	0.35092482	0	16.99572153
R. imeriensis	2.04857662	0.07154217	0.85552084	1.93441796	2.25527688	3.51544848	0.28719807	0.1606868	0.69884863	0.01111734	22.74464592
R. inundata	1.80056307	0.11360588	2.085254	0.15981462	3.07321944	1.7906451	0.26553121	0	0.50423881	0	32.47279883
R. latifolia	0.4516099	0.02041944	0.48019712	0.73112932	2.60041542	1.19771348	0.12864246	0.13556238	0.55506839	0.01009628	22.89245876
R. longisettis	1.30771996	0.08506943	3.71564744	0	8.01111954	4.49693359	0.12457866	0	1.4668247	0	20.34476254
R. macra	3.64379818	0.19394706	0.95525554	0.6737162	1.13920953	0.84135498	0	0	0.58653179	0.03895256	18.95786597
R. macrostachya	7.03856344	0.11186162	2.86126352	2.05119036	4.44092096	3.18591342	0.47522716	0	1.52867691	0	11.88998846
R. marisculus	4.20824418	0.03402091	1.82690989	0.0741216	0.72517573	1.35514466	0	0.30631754	0.4766808	0	24.23155433
R. megalocarpa	3.43772742	0.26284981	1.38826736	0.44953419	0.50727979	0.18747542	0	0.07605242	0.33958025	0.01220455	19.29177459
R. mixta	3.28904995	0.07662939	6.31581398	0	0.10786315	0.57074348	0.01765386	0	0.06964545	0	22.45066619
R. nitens	2.11977682	0.03418995	4.73353946	0.21545563	1.08199398	1.63610694	0	0	0.31743599	0	51.91832139
R. nivea	1.65347649	0.0511536	1.76803311	0.86490745	5.95273094	1.3140204	0.29898398	0.37718432	1.28050597	0	19.91933018
R. pallida	2.80306059	0.14698082	2.12623854	0	0.84591301	0.93595737	0	0.05705101	0.0106541	0.03585736	23.05032747
R. panicoides	7.04460899	0.05286449	0.98613886	0.48741725	2.80314793	4.40803407	0.22608713	0.08179034	0.94790354	0.01828646	17.20589555

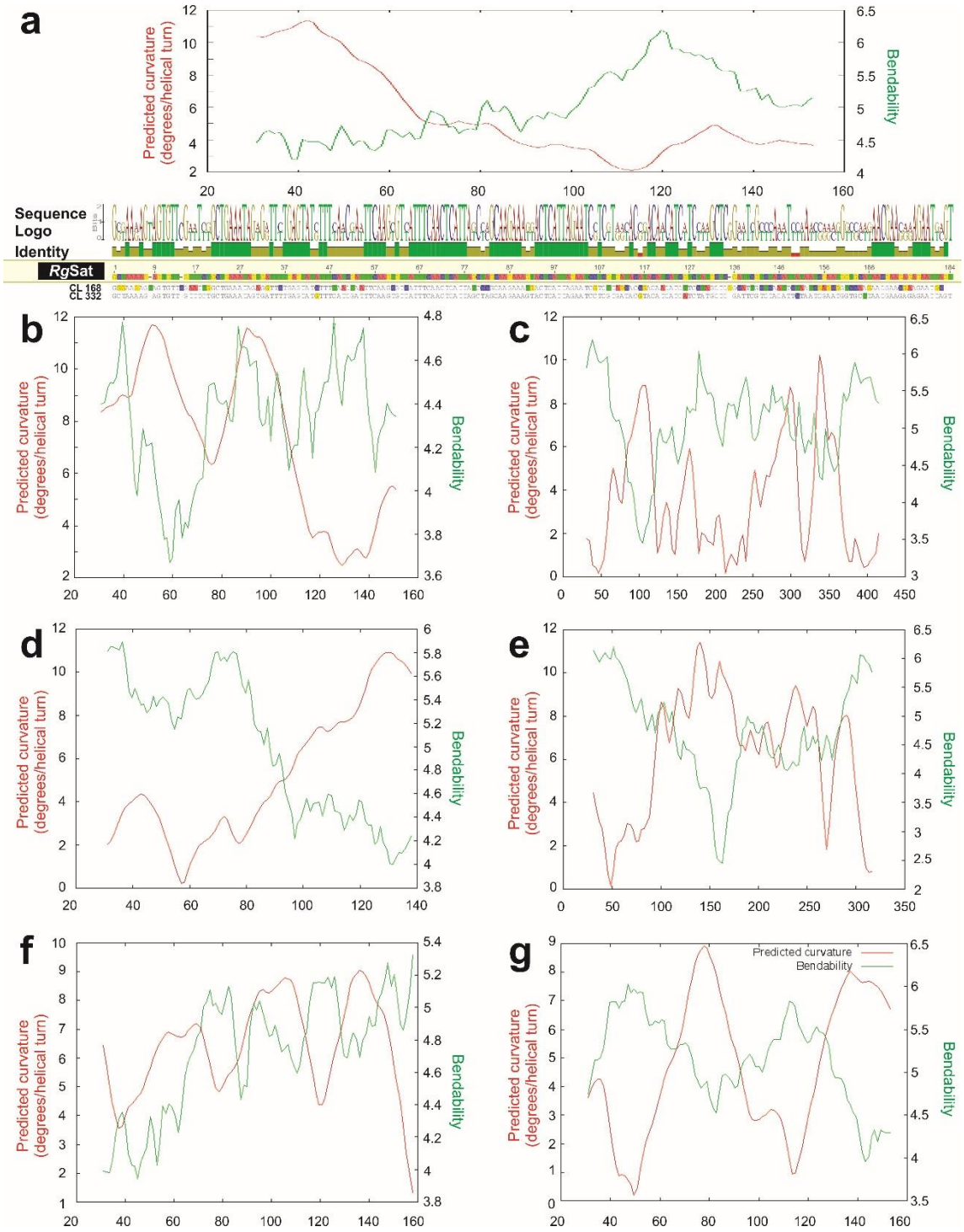
R. perplexa	4.27911074	0.01361338	0.82521372	0.15919853	0.38386268	0.52632629	0	0.01057855	0.06902071	0.02228433	17.70294561
R. pilulifera	1.98701159	0.07178162	0.62812681	1.82870076	3.17742777	2.01771065	0.65777247	0.17599289	0.80186241	0	13.99711519
R. pilulifera 2	3.64514266	0.39891057	0.45126217	1.30121858	3.11145916	0.96461075	0.13390761	0.12806673	1.04962802	0	19.57106298
R. polyphylla	0.26458408	0.27728603	0.81460261	2.04828979	7.75841684	11.414023	0	0.13788409	3.93688647	0.04034268	14.53470909
R. pubera	1.16753717	0	1.98851116	0.66770952	7.52839484	6.50823156	0.55273389	0.59342261	1.00620039	0	19.0862819
R. punctata	3.28144071	0.03795485	2.11176994	0.04756168	0.12567622	0.18284472	0	0	0.04315198	0.14961455	21.26243179
R. racemosa	0.67324209	0.09460816	3.87714945	1.04947773	5.2701825	7.94365253	0.01482973	0	2.83288959	0.01716403	22.04438757
R. radicans	3.61293165	0.11382203	0.5231062	1.07461391	6.64195843	1.57007452	0.90787651	0.04729986	1.74015963	0	15.09405436
R. rariflora	2.54150108	0.06282234	2.18653771	0	0.27978254	0.09703705	0	0.01315951	0.04748865	0	19.73994528
R. recognita	2.99543717	0.06211894	2.96094572	0	0.16078853	0.82401978	0	0.01269835	0.01870432	0.1928776	26.44053312
R. riedeliana	0.59988641	0.1767409	1.24528856	1.73404665	2.55682518	5.15487666	0.27325334	0.0405114	3.03088795	0.02581609	16.60951382
R. riparia	4.265976	0.05164682	2.51202962	0.33570431	0.75808324	0.80461651	0.09434493	0.09102112	0.13141814	0.02352231	18.67339268
R. robusta	5.28254792	0.01460117	0.82212246	2.5957798	11.3386502	9.18951237	0.03535019	1.82852699	1.34115546	0.01275681	19.46489035
R. rubra	1.26213616	0.14241572	1.43056895	0.32194583	8.83150063	7.67022685	0.43785126	0	2.54560385	0	23.79045284
R. schiedeana	1.17642657	0.09103796	1.38289013	2.92956544	10.5845136	16.2011537	0	0.27242104	4.3504227	0.0171822	18.00376069
R. scirpoides	2.44175135	0.0871508	2.5189639	0.89042889	1.16984659	3.18807563	0	0.08466624	0.25686551	0	37.62697547
R. sesleroides	0.43989855	0.0105674	0.95155709	1.36565154	5.3069214	2.82088019	0.48032498	0.08662807	1.01938503	0	29.90720352
R. sierrensis	2.98530995	0.03048165	0.84686438	0.35915795	2.34298796	3.69153829	0.03289916	0.09060408	5.40481736	0.06726985	19.33009737
R. tenuis	0.6605951	0.05716047	2.0421121	0.04449518	0.42578721	0.21897628	0.21980952	0.02466399	0.38995775	0	32.45931691
R. tenuiflora	3.56662746	0.04888407	0.81880815	0.43690136	2.00195536	1.07430378	0.6530606	0.14359695	0.49647882	0	22.37706421
R. tenuiflora 2	2.97877325	0.02597323	0.39927479	0	1.20088003	0.29028907	0.2918169	0.14718165	0.32033653	0	18.26427509
R. wrightiana	3.13159353	0.02212745	0.94566928	0	1.22662089	1.41436899	0	0.01050495	0.03397346	0	35.53244466



Supplementary Figure 1 – Classification of satDNA family *Tyba* into subfamilies (uncovered in the species comparative clustering analysis) and variants (based on the *Tyba*-like reads clustering analysis). Approximately-Maximum-Likelihood (AML) trees show the phylogenetic relationships based on consensus sequence similarity. Colored rectangles reflect the similarities between consensus sequences obtained in both analysis and are used to delimit sequences into subfamilies A (blue), B (yellow) and C (red). The clusters TL-CL01, 04, 05 and 12 formed a monophyletic clade and were all annotated as being similar to subfamily Tyba A (blue), being considered sequence variants A1-A4. Similarly, TL-CL03, 06, 07, 10 and CL11 formed a monophyletic clade that was annotated as being similar to Tyba C (red), thus being considered variants C1-C5. Although TL-CL02, 08 and 09 were annotated as Tyba B, only TL-CL02 and 08 were monophyletic, with TL-CL09 being in a polytomy with TLCL02+08 and the clusters annotated as Tyba A. However, the most similar satellite to TL-CL09 is TL-CL02 (75% pairwise identity), which is another evidence that it should be considered a variant of Tyba B (yellow). Thus, CL02, CL08 and CL09 were considered to be variants B1-B3



Supplementary Figure 2 – Dotplot of shared satDNAs found in the comparative clustering analysis of *Rhynchospora* species.



Supplementary Figure 3 - Bendability (green)/Curvature (red) propensity plots of shared SatDNAs *Rg*Sat (a), CL06 (b), CL10 (c), CL165 (d), CL239 (e), CL261 (f) and CL319 (g). Sequence logos, identity plot and alignment of satDNA clusters identified as *Rg*Sat are also presented at (a). The height of the nucleotides in the sequence logos is proportional to the number of times that base appear in the alignment. Green values on the identity plot represent high sequence identity, while yellow represent moderate and red represent low sequence identity. Base pairs that are different from the consensus sequence are colored in the alignment.

3.2 WHAT DRIVES LTR-RETROTRANSPOSON EVOLUTION IN *RHYNCHOSPORA* (CYPERACEAE) GENOMES?

*Artigo a ser submetido à revista Perspectives in Plant Ecology, Evolution and Systematics (https://www.journals.elsevier.com/perspectives-in-plant-ecology-evolution-and-systematics)

WHAT DRIVES LTR-RETROTRANSPOSON EVOLUTION IN *RHYNCHOSPORA* (CYPERACEAE) GENOMES?

Lucas Costa[1], Natália Castro[2], Chris Buddenhagen[2], André Marques[3], Andrea Pedrosa-Harand[1], Gustavo Souza[1]

[1] *Laboratory of Plant Cytogenetics and Evolution, Department of Botany, Federal University of Pernambuco, Recife-PE, Brazil*

[2] *AgResearch, Plant Functional Biology, Ruakura, New Zealand*

[3] *Max Planck Institute for Plant Breeding Research, Cologne, Germany*

ABSTRACT

Transposable elements (TEs), particularly LTR Retrotransposons (LTR-RTs), are the main components of most plant genomes. Because TEs can proliferate and move throughout host genomes, altering the genetic, epigenetic and nucleotypic landscape, they have been recognized as a relevant evolutionary force. Many factors have been shown to affect LTR-RT activity, ranging from genomic shock caused by polyploidy and chromosomal rearrangements to environmental stress such as heat and UV incidence. Here, we investigate LTR-RT evolution in the cosmopolitan, holocentric genus *Rhynchospora* Vahl. (Cyperaceae) by investigating whether factors such as the abundance of centromeric satDNA *Tyba*, *tempo*, phylogenetic relationships and environmental variables influenced LTR-RT abundance. We found that LTR-RT and *Tyba* abundance are inversely correlated, as their patterns show a phylogenetically significant contrasting pattern. Moreover, a decrease in LTR-RT content was observed in *Rhynchospora* lineages that have dispersed to low-precipitation environments in the northern hemisphere. In contrast, lineages that continued in rainier environments in South America showed bursts of LTR-RT insertions throughout the years, with evidence of purifying selection of TEs. A multivariate model showed that the studied traits presented stronger influence on LTR-RT abundance when associated with phylogenetic information, denoting that LTR

evolution is strongly correlated with species diversification. Altogether, our results present evidence of multi-trait influence on LTR-RT dynamics and provide a broader understanding of TE evolution in a macroevolutionary context.

Keywords: Transposable elements, Satellite DNA, genome evolution, genome ecology, holocentric chromosomes

INTRODUCTION

Transposable elements (TEs) and satellite DNA (satDNA) are the major components of the eukaryotic repeatome (Biscotti et al., 2015). While satDNAs are typically tandemly arranged, TEs constitute the repetitive fractions that can move through the genome using a DNA intermediate (DNA transposons) or an RNA intermediate (retrotransposons), varying in structure, abundance and transposition mechanisms (Neummann et al. 2019). High- throughput sequencing (e.g. Illumina) technologies has revolutionized the understanding of the origin and evolution of genomic repetitive fractions (Mehrotra and Goyal, 2014). It is now possible to identify, quantify and characterize repetitive DNA families by analyzing low-coverage genome sequencing or the byproduct of targeted capture sequencing (Costa et al., 2021; Dodsworth, 2015). Although previously discarded as "junk DNA", genomic studies have shown that repetitive DNA is not only a substantial part of the genome, but also have an active role in genome evolution (Gemmell, 2021). For example, long arrays of specific satDNAs have often been found in centromeric and telomeric regions, denoting a possible structural and/or maintenance function of these regions (Marques et al., 2015; Navajas-Pérez et al., 2009; Yu et al., 2017). On the other hand, the ability of TEs to move across the genome has been shown to directly impact speciation (Galindo-González et al., 2017). As TE copies can be inserted in any part of the genome, including inside or in the proximity of genes, they have the potential to alter the genetic and epigenetic landscape (Bennetzen and Wang, 2014; Galindo-González et al., 2017). Because of these widespread and abundant insertions, TEs are recognized as an evolutionary force that shapes genome structure through mechanisms of recombination and genomic rearrangements, as well as, structural functions, most notably in the centromeres (Arkhipova et al., 2017; Belyayev, 2014; Hartley and O'Neill, 2019).

Although repetitive DNA has been a focal point of genomic studies in all eukaryotic groups, plants in particular have shown a puzzling variability in repeat content (Ávila Robledillo et al., 2018; Ribeiro et al., 2020). The majority of the repetitive content in plants is comprised of Long Terminal Repeat Retrotransposons (LTR-RTs), a particularly large and diverse group of TEs (Bennetzen and Wang, 2014; Neumann et al., 2019). The Ty1-copia and the Ty3-gypsy elements are the two main superfamilies of LTR retrotransposons (Wicker et al., 2007). Since the earliest studies about TEs, it has been hypothesized that bursts of transpositions could appear as a response to stressful conditions (McClintock, 1984). The activation of TEs in plants has been shown to happen in response to various stimuli, including pathogen attack, tissue culture and environmental stresses (Grandbastien et al., 2005; Kraitshtein et al., 2010;

Liu et al., 2004; Matsunaga et al., 2015; Zhu et al., 2016). Polyploidization, and the resulting genomic shock, is another condition that can alter epigenetic patterns, triggering the mobilization of TEs (Galindo-González et al., 2017). Indeed, sudden bursts of activity and re-insertion of TEs after polyploidy events have been widely reported (Petit et al., 2010; Piegu et al., 2006; Senerchia et al., 2014).

The relationships between TEs and satDNAs are still intriguing. Extensive and complex links between these two large fractions exist in eukaryotic genome, creating a complex network of sequences that has a crucial effect on genome structure, function and evolution (Belyayev et al., 2019; Meštrović et al., 2015). Satellites and TEs often co-occur at heterochromatic regions, and, specifically in plant centromeres (Mata-Sucre et al., 2020; Plohl et al., 2014). Co-localization of satellites and LTR-RTs of the chromovirus family have been extensively reported (Cheng et al., 2002; Houben et al., 2007; Marques et al., 2015; Zhong et al., 2002). In addition, there is growing evidence of the involvement of TEs in generating a library of tandem repeats that can be dispersed throughout the genome and, in some cases, amplified into long arrays of new satDNAs (Ahmed and Liang, 2012; Belyayev et al., 2019; Kapitonov and Jurka, 2008). Lastly, the mobility of some TEs can serve as a dynamic system to "carry" satellite DNAs to different genomic regions, effectively helping its propagation (Vojvoda Zeljko et al., 2020).

In addition to the well-documented impact of genomic stress over TE activity, the effect of environmental-related stress has also been highlighted (Galindo-González et al., 2017). Historically, studies about the impact of environmental changes on the genome have mostly focused on genome size variation (Cacho et al., 2021; Enke et al., 2011; Grotkopp et al., 2004; Souza et al., 2019). Thus, as TEs are the "main culprit" of the remarkable genome size variation on eukaryotes (Elliott and Gregory, 2015), it is possible that increases/decreases in genome size are caused by bursts of insertions/deletions of TEs as a response to stress (Lyu et al., 2018; Schley et al., 2021). In this context, the idea that invasion of new environments could lead to a reduction of TE "load" has become increasingly popular (Hu et al., 2011; Ibarra-Laclette et al., 2013; Kelley et al., 2014; Lyu et al., 2018). The impact of environmental stress on TE activity has also been investigated experimentally (Galindo-González et al., 2017).

The abundance of TEs, particularly LTR-RTs, coupled with the great diversity of morphological adaptations and ecological niches make plants an interesting model to investigate genome-environment interactions (Lyu et al., 2018; Schley et al., 2021). Here, we investigate these potential interactions in species of the genus *Rhynchospora* Vahl.

(Cyperaceae), a cosmopolitan genus with approx. 400 species for which a robust dated phylogeny is available (Buddenhagen, 2016; Silva Filho et al., 2021; Thomas, 2020). The genus appears to have originated in South America and ended up colonizing the northern hemisphere via long-distance dispersal events, being found in both wet forests and drier savanna regions (Buddenhagen, 2016). The section *Eurhynchospora* (Gale, 1944), one of the clades to make this transition, is also associated with the highest chromosome numbers of the genus (2n = 26 to 2n = 36, (Burchardt et al., 2020; Ribeiro et al., 2018), which may suggest that the dispersion events and habitat shifts could have bern accompanied by polyploidy. The origin and initial diversification of the genus date to the Oligocene, where climatic shifts to a drier world favored the expansion of savannas and grasslands (Besnard et al., 2009). In the midst of this, adaptations such as changes to a C_4 photosynthesis pathway and reversals to insect pollination systems were detrimental for the genus success in wet forests, seasonally dry savannas and rock outcrops (Besnard et al., 2009; Buddenhagen, 2016; Galindo da Costa et al., 2021).

Rhynchospora has also been focal point of numerous cytogenetic/cytogenomic studies as a result of its holocentric chromosomes in which the centromere is dispersed along sister chromatids (Burchardt et al., 2020; Luceño et al., 1998; Marques et al., 2015; Ribeiro et al., 2017; Vanzela and Guerra, 2000). Another interesting feature of the chromosomes of *Rhynchospora* is the abundant presence of the satDNA *Tyba* in the centromeric regions of most species of the genus, being the only centromeric satDNA described in holocentric plant genomes (Marques et al., 2015; Ribeiro et al., 2017; Costa et al., unpublished). While most satDNAs appears as blocks in the heterochromatic regions of the genome (Barros e Silva et al., 2010; Lower et al., 2018; Macas et al., 2015; Ribeiro et al., 2020), *Tyba* has a peculiar dispersed distribution throughout the holocentromere, and appears as dispersed dot-like signals in interphasic nuclei (Marques et al., 2015; Ribeiro et al., 2017). These dispersed patterns are more common to LTR-RT elements, which due of its mobility, may not be restricted to heterochromatic regions (Dolgin and Charlesworth, 2008). It has been proposed that different repeat lineages in a genome behave similar to different species in an ecosystem (Brookfield, 2005; Schley et al., 2021). This could imply that repeat loss or amplification would be influenced by the available resources in the host genome and whether the amplification of a given repeat lineage is detrimental to the host (Novák et al., 2020). Because *Tyba* appears to have a structural role on *Rhynchospora* holocentromeres (Costa et al., unpublished), we hypothesize that other repetitive fractions, such as LTR-RTs, may be more likely to be targeted for elimination than *Tyba*.

In this study, we used comparative phylogenetic methods to investigate the forces driving LTR-RT evolution in the holocentric genus *Rhynchospora*. We statistically tested the impact of phylogenetic relationships, abundance of holocentromere-specific dispersed satDNA *Tyba*, *tempo* and ecological variables in LTR evolution. The hypothesis that multiple factors affect the evolution of LTRs in *Rhynchospora* was also tested, with a strongly differentiated pattern in the *Eurhynchospora* clade, which experienced different biogeographic, ecological and karyotypic conditions. We anchored our findings with a diversification rate shift analysis over a robust dated phylogeny of the genus and estimated the age of bursts of TE insertions in a subset of our sampling. Our results allowed us to discuss the following questions: 1) Do *Tyba* and LTR-RT abundances share similar patterns in a phylogenetic/temporal context? 2) Can colonization of new ecological divergent environment by the *Eurhynchospora* clade affect LTR-RTs abundance? 3) How these traces (and their interaction) drove the LTR evolution in the genus *Rhynchospora*?

MATERIAL AND METHODS

Sequence data acquisition and filtering

To analyze LTR-RT dynamics in *Rhynchospora* Vahl, we used a target-capture dataset obtained from Buddenhagen (2016). As the off-target portion of a target-capture sequencing can be similar to a genome skimming, we excluded the enriched genes from the data by mapping the raw datasets to the sequences of the target-genes as described by Costa et al. (2021). After filtering and accessing the quality of the repeat annotation by comparing with the species with lower proportion of annotated repeats from Costa et al. (2021), we ended up with 77 accessions representing 69 *Rhynchospora* species for our repetitive DNA analysis (accounting for approx. 20% of the genus), with representatives from the main clades and from most of the genus geographic distribution (**Supplementary Table 1)**. We also included data from the following outgroups: *Carex* L. (six spp.), *Chorizandra* R. Br. (two spp.), *Exocarya sclerioides* Benth., *Hypolytrum nemorum* (Vahl) Spreng. and *Scirpodendron ghaeri* (Gaertn.) Merr. (**Supplementary Table 1**; Buddenhagen, 2016). All datasets used were deposited on GenBank under project number PRJNA672127.

Repetitive DNA analysis

The *off-target* datasets of the 77 *Rhynchospora* accessions were uploaded to the RepeatExplorer pipeline (Novak et al., 2013) hosted at the Galaxy online platform (https://repeatexplorer-elixir.cerit-sc.cz/). RepeatExplorer allows the identification of repetitive sequences by means of a graph-based clustering algorithm that group sequences based on similarity. These clusters are identified by cross-referencing against a platform of protein domains of known repetitive elements (Neumann et al., 2019; Novak et al., 2013). The uploaded *Rhynchospora* reads were filtered by quality using the default settings of 95% of bases with quality value equal or above the cut-off value of 10. Clustering analysis was also performed with default settings of 90% similarity over a minimum overlap of 55% sequence length. Concurrently, we employed the TAREAN (Tandem Repeat Analyzer) tool to check the clusters for predictions of tandem arrangement, building consensus sequences for these clusters (Novák et al., 2017). To identify sequences similar to the holocentromere-specific satellite DNA *Tyba*, we uploaded a FASTA file containing the consensus sequences of previously described *Tyba* variants (Marques et al., 2015; Ribeiro et al., 2017) to be used as reference for cluster annotation.

Phylogenetic analysis and molecular dating

In order to investigate LTR-RT evolution in a phylogenetic context, we used the robust RaxML topology constructed by Buddenhagen (2016) based on the 256 target loci obtained by target-capture sequencing. This tree contained 115 *Rhynchospora* accessions, but as the reads of some of these were not sufficient for the RepeatExplorer analysis, we pruned the original tree leaving only 69 *Rhynchospora* species and the 11 outgroup species. We used the *drop.tip* function implemented in the package *phytools* (Revell, 2012) in Rstudio (R Core Team, 2019) to prune the tree. Divergence times for this topology were estimated on BEAST v.1.8.3 (Drummond and Rambaut, 2007) through CIPRES Science Gateway using the pruned tree as fixed topology. We used the same calibration points used by Buddenhagen (2016), following a normal distribution with a 10% standard deviation. Uncorrelated relaxed lognormal clock (Drummond and Rambaut, 2007) and Birth-Death speciation model (Gernhard, 2008) were applied. Two independent runs of 100,000,000 generations were performed, sampling every 10,000 generations. After removing 25% of samples as burn-in, the independent runs were combined and a maximum clade credibility (MCC) tree was constructed using TreeAnnotator v.1.8.2 (Rambaut and Drummond, 2013). In order to verify the effective sampling of all parameters and assess convergence of independent chains, we examined their posterior distributions in TRACER. The MCMC sampling was considered sufficient at effective sampling sizes (ESS) equal to or higher than 200.

Diversification rate shift analysis

We used the nuclear marker phylogeny to estimate diversification rates under an speciation/extinction model analysis implemented in the software BAMM (Rabosky, 2014). As we sampled only 69 out of 400 *Rhynchospora* species (Silva Filho et al., 2021), we had to account for phylogenetic incompleteness by informing the percentage of recorded species in each major clade. We manage this by first checking the number of species recorded for each of the traditional *Rhynchospora* sections created by Kükenthal (1939). Although these sections were not monophyletic in our *Rhynchospora* phlogeny, we could associate each of the sections to one of the five clades recovered in our phylogenetic analysis (**Supplementary Table 2**). Priors for the BAMM control file were generated using the dated phylogenetic tree input into the function set BAMM priors in the package BAMMtools v. 2.5.0 (Rabosky, 2014) implemented in R. The control file was set for 10,000,000 generations and the analysis was run twice as recommended, returning similar results. Resulting MCMC log likelihoods were tested against generation number using the CODA package (Plummer et al., 2006) implemented in R. All remaining outputs contained in the event data file were analyzed using BAMMtools. BAMMtools was then also used to produce a figure showing the best rate shift configuration as well as graphics of diversification through time for clades I, II+III, IV and V of our *Rhynchospora* phylogeny.

Environmental data acquisition

To gather environmental information of *Rhynchospora* species, occurrence data of the 69 species sampled here were downloaded from the Global Biodiversity Information Facility (GBIF) website (https://www.gbif. org). To minimize the effect of erroneous GBIF distribution data, we used the function CoordinateCleanner (Zizka et al., 2019) implemented in the R software (R Core Team, 2019). CoordinateCleanner allowed us to remove duplicate records, geographic outliers, oceanic points and zero coordinate points. From the collection coordinates of each species, we extracted values for the 19 climatic variables available in the WorldClim 1.4 (5 min) generic grid format (Hijmans et al., 2005) utilizing the package "raster" 2.6–7 (Hijmans & van Etten 2012) implemented in R. In order to have typical variable values for each species, we calculated species median for latitude and the 19 climatic variables.

Phylogenetic comparative methods and statistical analysis

We used the pruned dated tree to reconstruct the ancestral states of LTR-RT abundance, *Tyba* abundance and the median values of the environmental variables across *Rhynchospora*

phylogeny. For this, we used the *FastAnc* function of the phytools package (Revell, 2012) implemented in R and plotted the results with the *contMap* function, also in phytools. The phylogenetic signal of these traits was assessed using Pagel's lambda (λ) via the *phylosig* function implemented in *phytools*.

In order to investigate whether genomic (*Tyba* abundance), environmental (climatic variables), temporal and/or phylogenetic factors were contributing to LTR-RT content evolution in *Rhynchospora*, we employed a series of linear regression analyses using LTR-RT abundance as the response variable. For *Tyba*, we simply used the genomic abundance of the clusters identified as *Tyba* for each species as the explanatory variable. For the climatic variables, we first assessed correlations between them with Pearson's correlation coefficient, to avoid the selection of highly correlated variables ($r>0.75$). The median values of the selected variables were used as explanatory variables. The temporal factor was explored by using the divergence age (the age of the most recent ancestor) for each species as the explanatory variable. To be able to use the phylogenetic information as an explanatory variable in the regression analyses, we employed a Phylogenetic Eigenvector Regression (PVR) approach, in which species-specific eigenvectors are extracted from a phylogenetic distance matrix (Diniz-Filho et al., 1998; Guénard et al., 2013). This was achievable by using the *PVRdecomp* function built in the PVR package (Santos 2018) implemented in R. After the selection of the explanatory variables, we conducted simple linear regressions (SLR) between LTR-RT abundance and each variable separately, also using the software R.

We also conducted a multiple linear regression (MLR) analysis in R using *Tyba* abundance, phylogenetic eigenvectors, time of divergence and the strongest-correlated environmental variable Precipitation of Wettest Month (Bio13) as predictors for LTR-RT abundance. In addition to the MLR, we conducted a regression commonality analysis (CA) using the *commonalityCoefficients* function of the package *yhat* (Nimon et al. 2021) implemented in R. This analysis decomposes the R^2 variance of a MLR model into unique and shared effects of the predictors, effectively showing the contribution of each variable and group of variables to the strength of the model (Ray-Mukherjee et al., 2014). Graphs obtained in R were further edited in CorelDraw X7.

Temporal LTR-RT dynamics

To investigate the temporal LTR-RT dynamics in *Rhynchospora*, we calculated the distributions of pairwise divergence between Illumina reads mapped to the reverse transcriptase

(RT) domain of abundant LTR-RT families (Usai et al., 2017; Mascagni et al., 2020). First, we selected four species from our sampling, of which two presented low LTR-RT content [*R. alba* (clade I, 1.60% LTR-RT content) and *R. macrostachya* (clade V, 7.63% LTR-RT content)] while the other two presented high LTR-RT content [*R. robusta* (clade IV, 20.53% LTR-RT) and *R. barbata* (clade V, 26.41%)]. We extracted the reverse-transcriptase sequence of the most abundant LTR-RT element of each of these species (Ty1-copia/Angela for *R. alba*, *R. macrostachya*, and *R. robusta* and Ty3-gypsy/Athila for *R. barbata*) using the Protein Domains Filter function of RepeatExplorer (Neumann et al., 2019). Then, we mapped the raw reads of the four species to the corresponding reverse-transcriptase domain sequence using the Low Sensitivity preset (5 iterations) of the Geneious Read Mapper v. 6.0.3 plugin implemented in Geneious v. 7.1.9 (Kearse et al., 2012). From the mapped reads, we randomly selected 100 reads and calculated pairwise divergence using the MAFFT plugin (Katoh, 2002) under the Kimura two-parameter model of sequence evolution (Kimura, 1980). The pairwise distances were then converted to millions of years using the rice substitution rate of 4.9×10^{-9} substitutions/site/year (Mascagni et al., 2020; Usai et al., 2017). The resulting values were used to build frequency histograms, where peaks indicate bursts of LTR-RT insertion.

RESULTS

Phylogenetic framework

Our phylogenetic tree of 69 *Rhynchospora* species was identical to previously reported topologies that used the same genomic dataset (Buddenhagen, 2016, Costa et al., unpublished). Briefly, the genus was shown to have originated approx. 37.8 Mya, being further divided into two main lineages (**Figure 1, Supplementary Figure 1**). One of these lineages diversified into clade V, while the other lineage further diversified into four different clades (I-IV). Although the lineage that originated clade I split from clades II-IV approx. 31.8 Mya, the crown node of the clade I is relatively recent, at approx. 10.4 Mya. The lineage that originated clade IV diverged from the lineage that originated clades II + III at approx. 30 Mya, while the latter clades separated at approx. 27 Mya. Despite the recent age of its most recent common ancestor (MRCA), clade I was the most species-rich clade in *Rhynchospora* (**Supplementary Table 2**). The 95% credible set of rate shift from our BAMM analysis detected a single shift in the diversification rate for *Rhynchospora* at the base of this clade (**Figure 1**). Density plots of

speciation through time illustrated the steady increase of diversification for clade I at its first split, contrasting to a stable rate for clades II-V (**Figure 1**).

Variation of LTR-RTs and Tyba *abundances across* Rhynchospora *phylogeny*

Our RepeatExplorer analysis showed total abundance of LTR retrotransposons varying from 0.23% of the genome in *R. decurrens* to 26.79% in *R. schiedeana* with mean value = 6.02% and median value = 3.80% across the genus, which are generally lower values, showing that LTR are not the most abundant fraction in most *Rhynchospora* genomes (**Figure 1, Supplementary Table 1**). Generally, LTR-RT abundance seemed to have clade/subclade specific patterns. Most notably, LTR-RT abundance in clade I were generally low, varying from 0.23% in *R. decurrens* to 4.93% in *R. fernaldii* (mean = 1.50%, median = 1.54%). In contrast, LTR-RT abundances in the sister group to clade I (formed by clades II, III and IV) ranged from 0.96% in *R. cephalotes* to 26.79% in *R. schiedeana* (mean = 7.20%, median = 5.20%). Within clade V, two subclades presented different LTR-RT abundance dynamics. The subclade with a most recent crown node (approx. 5.5 Mya) showed overall lower LTR-RT content (mean = 4.09%, median = 4.18%), the subclade with older crown node (approx. 24 Mya) presented higher abundances of LTR-RT (mean = 14.67%, median = 16.5%). The phylogenetic dependence of LTR-RT abundance was further confirmed by the high phylogenetic signal (λ = 0.99) for this trait across our *Rhynchospora* phylogeny.

Our repetitive DNA analysis confirmed previous reports (Costa et al. unpublished) that showed that species from clade V did not present the holocentromere-specific satellite DNA *Tyba*-like. Thus, the abundance of *Tyba* varied from zero (clade V) to 4.18% in *R. glaziovii* (clade II). Once again, clade I stood out by showing overall higher genomic proportion of *Tyba* (mean = 2.51%, median = 2.56%) than its sister lineage of clades II to IV (mean = 1.27%, median = 1.31%). These clade-specific patterns were reflected in the high phylogenetic signal (λ = 0.9) for *Tyba* abundance (**Supplementary Table 1**).

Factors influencing LTR-RT abundance in Rhynchospora

We thus investigated the factors affecting LTR-RT abundance in *Rhynchospora* genomes by linear regressions. For all the regression models, the use of the untransformed raw data lead to high non-normality and heteroscedasticity of the residuals. Transformation of the

response variable (LTR-RT abundance) to its log(10) values helped to improve the model assumptions. Thus, all regression models here reported were evaluating the ability of the response variables to predict LTR-RT abundance at a logarithmic scale. The predictor variable with the strongest effect on LTR-RT abundance was the phylogenetic eigenvectors, which represented phylogenetic relatedness (**Table 1, Figure 2A**). This result corroborated the high phylogenetic signal observed for LTR-RT abundance, as phylogenetically related species tend to have similar LTR-RT content (**Supplementary Figure 2**). Similarly, LTR-RT abundance was shown to increase depending on temporal factors (**Table 1**, **Figure 2B**). This would mean that LTR-RT content of species with older diversification events tend to be higher than that of species with more recent diversification events. We also used the abundance of the holocentromere-specific satDNA *Tyba* as a predictor to LTR-RT abundance. Our results showed that as *Tyba* content increases, LTR-RT content is predicted to decrease in *Rhynchospora* genomes (**Table 1, Figure 2C**).

To check for the influence of environmental factors over LTR-RT abundance, we downloaded a total of 54,510 occurrence points and extracted values of 19 bioclimatic variables and latitude. After the exclusion of highly autocorrelated variables, we performed simple linear regressions using seven bioclimatic variables and latitude (**Table 1).** Although statistically significant, Latitude and Precipitation of Driest Month (Bio14) presented low R^2 values (0.13 and 0.07 respectively), showing poor fitting. For the remaining significant models using environmental variables, Precipitation of Wettest Month (Bio13) was shown to be the strongest predictor (R^2 = 0.22, F = 19.86), showing that *Rhynchospora* species occupying rainier environments tend to have larger amounts of LTR-RTs than species distributed in areas with less rain (**Figure 2D**). Median values for Bio13 ranged from 85 mm in *R. alba* to 402 mm in *R. polyphilla* (mean = ~224.43 mm, median = 205 mm) and were shown to have a moderately high phylogenetic signal (λ = 0.65) when plotted across our *Rhynchospora* phylogeny (**Supplementary Figure 1**).

After the simple linear regressions, we designed a multiple linear regression (MLR) model using *Tyba* abundance, age, phylogenetic eingenvectors and Bio13 as predictor variables. The overall model was statistically significant ($p < 0.001$, F = 21.43) and explained 54.6% of the variation on LTR-RT abundance. However, only phylogenetic eigenvectors and age of MRA were shown as contributing significantly to the model ($p < 0.001$), while Bio 13 and *Tyba* abundance were not significant ($p = 0.09$ and $p = 0.17$, respectively). As collinearity can influence *p*-value statistics of predictor variables in a multivariate model, we undertook a

commonality analysis on our MLR analysis in order to decompose the variance explained by the model among the predictor variables, effectively showing the contribution of each individual and shared set of predictors. This analysis showed that the most important effect on LTR-RT abundance is the combination of *Tyba* abundance with the phylogenetic information (represented by the phylogenetic eigenvectors), accounting for approx. 14% of the variation (**Figure 3**). The unique contribution of phylogenetic information was the second most contributing factor, accounting for approx. 10% of the variance of LTR-RT abundance. This represents an additional evidence that closely related species tend to have similar LTR-RT content, with *Tyba* content also varying depending on phylogenetic positioning. Other important contributing factors to LTR-RT abundance were the combining effect of *Tyba* abundance, phylogenetic information and Bio 13 (5.2% of the variance) and the combining of all the predictor variables (7.9% of the variance) (**Figure 3**). These results show that the explanatory power of *Tyba* abundance, time (age of MRA) and environment (Bio13) is highly dependent of the phylogenetic relationships, as these predictors individually are insufficient to explain a large portion of LTR abundance variation.

Time of LTR-RT insertion events

We investigated the dynamics of LTR-RT insertion events in the genome of four *Rhynchospora* species by converting the pairwise distance between aligned reads (mapped to reverse transcriptase domains of abundant LTR-RTs) to millions of years ago. We selected two species with larger amounts of LTR-RTs and two species with smaller amounts. Species with a large LTR-RT content such as *R. robusta* (20.53%) and *R. barbata* (26.41%) were shown to have constant bursts of insertions in the last 60 My, with peaks of insertion at around ~28 and ~17 Mya respectively (**Figure 4**). On the other hand, species with low LTR-RT content from the recently diversified clade I [*R. alba* (1.60%)] and from the most recently-diversified subclade of clade V [*R. macrostachya* (7.63%)] showed peaks of insertion at approx. 5 Mya (**Figure 4**).

DISCUSSION

Tyba and LTRs follow contrasting evolutionary paths

Although they have different composition, genomic organization, function, proliferation mechanisms etc., we demonstrated that the abundance of centromeric satDNA *Tyba* and LTRs is significantly correlated in *Rhynchospora*. The satDNA *Tyba* stands out not only for being the only centromere-specific satellite found in a holocentric plant species so far (Marques et al., 2015), but also for its remarkable phylogenetic reach, being present in four out of five major *Rhynchospora* clades (Costa et al. unpublished). Both LTRs-RT and *Tyba* traits presented moderate-high phylogenetic signal and clearly contrasting patterns throughout the phylogeny, where species with larger amounts of *Tyba* tended to have less LTR-RT content than species with lower content of this satDNA. In addition to this, *Tyba* abundance was shown to be a significant, albeit not strong, predictor to LTR-RT abundance. Due to its association with the holocentromere, *Tyba* presents a unique case of a dispersed satDNA distribution, which is a pattern more common to TEs, especially LTR-RTs (Dolgin and Charlesworth, 2008). There has been proposed that the genomic abundance of different types of repetitive elements may be constrained by a "carrying capacity" of a given genome, with replication of certain repeats being reduced if the host exceeds optimal values of genome size (Schley et al., 2021). In a scenario where *Tyba* presents structural advantages for the centromeres of *Rhynchospora* species (Costa et al., unpublished), it may be possible that it escapes similar constraints, while other dispersed repeats such as LTR-RTs may be selected for elimination, presenting an interesting dynamic between satDNA and TE abundance in the genus.

Our commonality analysis showed that, in a multiple linear regression model, *Tyba* is a stronger predictor of LTR-RT abundance when combined with phylogenetic information than on its own. In addition to this, LTR-RT abundance patterns, much like *Tyba* patterns (Costa et al. unpublished) seem to be highly clade/sub-clade specific. Thus, these contrasting patterns could be the result of different pressures applied to these genomic fractions by speciation events. Polyploidization, for example, is a genomic event that impacts both species diversification and repetitive DNA dynamics (Alix et al., 2017; Parisod et al., 2010). In our data, clade I presented the largest amounts of *Tyba* and the lowest abundances of LTR-RT of the genus, in addition to a steep increase in the speciation rate. This clade, commonly referred to as *Eurynchosporae* (Gale, 1944) is karyotypically characterized by high chromosome numbers (2n = 26 to 2n = 36) assumed to be related to higher ploidy levels (3x, 4x) (Burchardt et al., 2020; Ribeiro et al., 2018). In a scenario in which polyploidy drove the increase of speciation in this clade, it is possible to assume that the genomic shock could quickly change the abundance dynamics between satDNA and TEs in comparison to the rest of the other

Rhynchospora clades. In the case of satDNA amplification after polyploidization events (*e.g.* Yang et al., 2018), a holocentric dispersed satDNA such as *Tyba* could quickly become the most abundant genomic component. As for LTR-RTs, most studies with synthetic polyploids reveal a tendency of bursts of TE insertions after genome duplication (Madlung et al., 2004; Petit et al., 2010; Yaakov and Kashkush, 2012). However, it has also been shown that genome downsizing after duplication [related to diploidization mechanisms], often involves the loss of large amounts of repetitive DNA (Renny-Byfield et al., 2013).

Eurhynchosporae (clade I) appears to have a much smaller genome size per monoploid complement (1C*x* = 0.07 pg in contrast to 1C*x* = 0.36 pg in the rest of the genus), indicating a possibly efficient diploidization mechanism (Burchardt et al., 2020; Ribeiro et al., 2018). Thus, the polyploidization event of clade I could have impacted LTR-RT and *Tyba* abundance patterns in a contrasting manner, leading to an indirect relationship between these traits in a phylogenetic framework. Interestingly, species from clade V, which is formed by species without *Tyba* (Ribeiro et al., 2017; Costa et al., unpublished), seems to have an ample range of LTR-RT abundances. As expected, this variation also appears to follow a phylogenetic significant pattern, as one of the subclades have overall lower LTR-RT content than the other. Similarly to what happened in clade I, the subclade with less LTR-RTs might also have undergone a polyploidy in its origin (~5.5 Mya), going from $n = 5$ to $n = 9$ (Ribeiro et al., 2018). Thus, even in the absence of *Tyba*, LTR-RT abundance could be following a pattern linked to polyploidy events, in which recent polyploid lineages tend to have less LTR-RTs than diploid species due to diploidization. Alternatively, it is possible that the high chromosome numbers in clade I (*Eurynchosporae*) are a result of agmatoploidy, since chromosomal fissions are highly frequent in holocentric plants (Márquez-Corro et al., 2019). This hypothesis is supported by genomic analyses, which did not detect whole genome duplication in *R. alba* (clade I, 2n = 26) (Marques, unpublished). In this scenario, it would be chromosomal rearrangements, probably linked to the reduction of LTRs and genome downsizing, that generated the discrepant evolutionary pattern observed here in this clade.

Dispersion of Rhynchospora *species to new environments was accompanied by loss of LTR, and amplification of* Tyba

We found that tempo of evolution, phylogenetic relationships and environmental factors seem to also contribute to predict LTR-RT abundance in *Rhynchospora*. The genus probably originated in South America, with posterior long-distance dispersal to the northern hemisphere, which became its center of diversity (Buddenhagen, 2016; Spalink et al., 2016). These recent northern-hemisphere lineages appear at our phylogeny mostly as species with lower LTR-RT content habiting comparatively less rainy environments, such as the clade I (*Eurhynchosporae*) and the subclade with lower LTR-RT abundance from clade V. Our LTR-RT insertion date analysis also imply that most of the LTR-RT content in these species might have been lost throughout the years, as most of the few elements still present in these genomes seem to have fairly recent insertion bursts. Comparatively, older lineages were shown to predominantly stay at rainier environments and our temporal analysis of LTR-RTs showed constant insertion activity for long periods, possibly leading to the observed high LTR-RT abundances. Genome-environment interactions have been extensively studied using the genome size as a response variable to a series of environmental stress-inducing factors, with some of these papers showing a decrease of DNA content associated with dry environments (Faizullah et al., 2021; Souza et al., 2019; Trávníček et al., 2019).

The growing availability of NGS data is allowing the investigation of these interactions from the point of view of the repetitive DNA, especially LTR-RTs, as these represent the most dominant fraction on the majority of eukaryotic genomes (Schley et al., 2021). In *Rhynchospora*, precipitation seems to be an important factor regarding LTR-RTs abundance, with lineages that stabilize in high-precipitation environments maintaining constant insertion activity throughout the years leading to high LTR-RT content. In contrast, species that migrated to drier, low-precipitation environments seem to have a steep decrease in LTR-RT content. The concept of environmental stress affecting TEs activity has been investigated since the earliest studies concerning these elements (McClintock, 1984). Since then, the advancement in sequencing techniques have been essential for theoretical and experimental investigations of the genome-environment interaction. It has already been shown that effects on transposition activity can be induced by environmental stress caused by factors such as heat (Cao et al., 2014; Matsunaga et al., 2015), UV incidence (Kimura et al., 2001; Ramallo et al., 2008) and water availability (Kalendar et al., 2000). While TEs can affect the genetic and epigenetic landscape of a genome, they can also indirectly cause nucleotypic effects by contributing to genome size changes, which in turn can affect cell size, division rate and physiology (Hidalgo et al., 2017). In palms, for example, it was proposed that TE dynamics could have played a role in adaptation

to different environments by affecting both epigenetic (amplification of TEs related to water stress-response genes) and nucleotypic factors (lower TE abundance in arid-adapted palm species) (Schley et al., 2021). As *Rhynchospora* began to diversify in the Oligocene, a period in which low levels of CO_2 increased aridity and plant water demand (Osborne and Sack, 2012), loss of LTR-RT could have become a common strategy for nutrient conservation as a response to environmental stress.

While the direction of TE evolution is still point of debate, a tendency of genome downsizing has been reported for "invasive" species, which could indicate TE loss when colonizing new environments (Pandit et al., 2014). A possible explanation of this pattern could rely on the concept of purifying selection of TEs, which postulates that as TEs proliferate uncontrollably, host fitness decreases, generating the need of efficient removal (Charlesworth and Charlesworth, 1983). This form of selection was observed in mangroves, which showed a significant reduction of TE content when compared with other angiosperm species directly related to the colonization of intertidal environments (Lyu et al., 2018). These "unloading" of TEs under environmental stress was also observed among populations of the same species, as maize landraces were also shown to suffer TE loss when inhabiting high altitudes (Bilinski et al., 2018; Díez et al., 2013). Thus, we favor a scenario in which as *Rhynchospora* species invaded new, less-rainy environments and began to rapidly diversify, TE elimination occurred in order to maintain species fitness. This is corroborated by our analysis of temporal dynamics of LTR-RTs, which showed very recent (post-migration) bursts of insertion for the most abundant elements of species with low LTR-RT abundance. Moreover, as some of these invasive lineages (such as *Eurhynchosporae*) potentially suffered polyploidy (or multiple chromosome fissions) events upon diversification in these new environments (Burchardt et al., 2020; Ribeiro et al., 2018), unequal recombination of TEs could have caused chromosomal aberrations and accelerated deleterious effects upon these elements (Petrov et al., 2011; Robberecht et al., 2013).

LTR abundance is mainly driven by complex traits interactions

Phylogenetic information is an important factor when performing statistical inferences in a group of species, as investigated traits should be checked for phylogenetic non-independence (Sakamoto and Venditti, 2018). Here we used an eigenvector-based approach in order to decompose phylogenetic distances into a series of vectors to be used in regression

analyses (Diniz-Filho et al., 1998; Guénard et al., 2013). Altogether, our results point to an important contribution of phylogenetic information to the LTR-RT abundance patterns in *Rhynchospora*, associated with factors such as *Tyba* abundance, precipitation and diversification time. Because they are in general scattered through the genome, repetitive elements, especially TEs, are often responsible for characteristic differences between chromosomes and chromosomal regions within and between species such as promoting expansion of heterochromatic regions, rearrangements, etc. (Brookfield, 2005). In addition to this, these repetitive elements are heavily involved in chromosomal rearrangements, which are a driving force of speciation, normally reflecting phylogenetic relationships (Chen et al., 2020; Kiazim et al., 2021; Oliveira da Silva et al., 2020). Lastly, repeat abundance and composition have both been shown to be useful in the reconstruction of phylogenetic relationships, producing topologies similar to traditional phylogenetic methods (Dodsworth, 2015; Vitales et al., 2019).

Given this, our results provide another evidence to how LTR-RT evolution is strongly correlated with species diversification. Not only were phylogenetic eigenvectors the strongest contributors to overall LTR-RT abundance variation, they were also helpful to provide insights in how the other studied traits may be affecting LTR-RT content in *Rhynchospora*, as these were stronger predictors when combined with phylogenetic information. Due to the apparent clade-specific biogeographic patterns observed in the genus (Buddenhagen, 2016), it makes sense that environmental factors such as precipitation contribute more to the overall model when combined with phylogenetic information. Similarly, we have already shown that characteristics such as sequence structure, GC content and monomer size of the satDNA *Tyba* are intimately linked to the phylogenetic relationships of *Rhynchospora* (Costa et al., unpublished). Here, *Tyba* abundance was shown to have moderate phylogenetic signal and presented a much stronger contribution to the overall LTR-RT abundance when combined with phylogenetic information than on its own. Lastly, time of diversification was a moderate predictor of LTR-RT abundance, with a stronger contribution when combined to all other predictors, along with phylogeny. This relationship between time and LTR-RT abundance seems to be directly related to the relatively recent increase of speciation rate observed in clade I, which possessed the lowest LTR-RT abundances in the genus (**Figure 5**). When studying TE content in a macroevolutionary context, a temporal framework of TE insertion may be necessary to understand whether changes in abundance happened over longer or shorter periods of time (Lyu et al., 2018; Mascagni et al., 2020; Usai et al., 2017). We provided evidence for

the importance of diversification times for LTR-RT abundance by showing that species with larger amounts of LTR (older lineages) seemed to have near constant rates of TE insertion throughout the years, while species with low LTR (recently diversified lineages) have high rates of TE removal and recent insertion bursts in the last 5 My.

CONCLUSION

Altogether, we favor a scenario in which an ancestral of *Rhynchospora* presented a moderately large number of LTR-RTs and inhabited high-precipitation environments. While lineages that stayed in these environments preserved a high LTR-RT content, lineages that rapidly diversified into less-rainy regions suffered a quick loss of LTR-RTs due to environmental constrains. In contrast, the dispersed satDNA *Tyba* presented contrasting patterns of abundance, representing most of the repetitive fraction of these low-LTR-RT genomes. We hypothesized that, as *Tyba* seems to have a functional role in the centromeres of *Rhynchospora* (Costa et al., unpublished), it could have been less likely to suffer from deleterious effects than LTR-RTs. As the importance of repetitive DNA to genome and species evolution is being reinterpreted (Gemmell, 2021), studies of this genomic fraction in a macroevolutionary context are even more required. With biodiversity projects generating an ever-increasing amount of genomic data, and given the possibilities to mine available data for repetitive DNA information (Costa et al., 2021), evolutionary patterns of LTR-RTs and other repetitive elements will become clearer. We believe that our findings not only help to elucidate what drives LTR-RT abundance in holocentric species, but also provide a broader understanding of repeat dynamics associated with phylogenetic, environmental and temporal factors.

REFERENCES

Ahmed, M., Liang, P., 2012. Transposable Elements Are a Significant Contributor to Tandem Repeats in the Human Genome. Comparative and Functional Genomics 2012, 1–7. https://doi.org/10.1155/2012/947089

Alix, K., Gérard, P.R., Schwarzacher, T., Heslop-Harrison, J.S. (Pat), 2017. Polyploidy and interspecific hybridization: partners for adaptation, speciation and evolution in plants. Annals of Botany 120, 183–194. https://doi.org/10.1093/aob/mcx079

Arkhipova, I.R., Yushenova, I.A., Rodriguez, F., 2017. Giant Reverse Transcriptase-Encoding Transposable Elements at Telomeres. Molecular Biology and Evolution 34, 2245–2257. https://doi.org/10.1093/molbev/msx159

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

Barros e Silva, A.E., Marques, A., dos Santos, K.G.B., Guerra, M., 2010. The evolution of CMA bands in Citrus and related genera. Chromosome Res 18, 503–514. https://doi.org/10.1007/s10577-010-9130-2

Belyayev, A., 2014. Bursts of transposable elements as an evolutionary driving force. J. Evol. Biol. 27, 2573–2584. https://doi.org/10.1111/jeb.12513

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. IJMS 20, 1201. https://doi.org/10.3390/ijms20051201

Bennetzen, J.L., Wang, H., 2014. The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes. Annu. Rev. Plant Biol. 65, 505–530. https://doi.org/10.1146/annurev-arplant-050213-035811

Besnard, G., Muasya, A.M., Russier, F., Roalson, E.H., Salamin, N., Christin, P.-A., 2009. Phylogenomics of C4 Photosynthesis in Sedges (Cyperaceae): Multiple Appearances and Genetic Convergence. Mol. Biol. Evol. 26, 1909–1919. https://doi.org/10.1093/molbev/msp103

Bilinski, P., Albert, P.S., Berg, J.J., Birchler, J.A., Grote, M.N., Lorant, A., Quezada, J., Swarts, K., Yang, J., Ross-Ibarra, J., 2018. Parallel altitudinal clines reveal trends in adaptive evolution of genome size in Zea mays. PLOS Genetics 14, e1007162. https://doi.org/10.1371/journal.pgen.1007162

Biscotti, M.A., Olmo, E., Heslop-Harrison, J.S., 2015. Repetitive DNA in eukaryotic genomes. Chromosome Res 23, 415–420. https://doi.org/10.1007/s10577-015-9499-z

Brookfield, J.F.Y., 2005. The ecology of the genome — mobile DNA elements and their hosts. Nat Rev Genet 6, 128–136. https://doi.org/10.1038/nrg1524

Buddenhagen, C.E., 2016. A view of Rhynchosporeae (Cyperaceae) diversification before and after the application of anchored phylogenomics across the angiosperms (PhD thesis). Florida State University, Florida, USA.

Burchardt, P., Buddenhagen, C.E., Gaeta, M.L., Souza, M.D., Marques, A., Vanzela, A.L.L., 2020. Holocentric Karyotype Evolution in Rhynchospora Is Marked by Intense Numerical, Structural, and Genome Size Changes. Front. Plant Sci. 11, 536507. https://doi.org/10.3389/fpls.2020.536507

Cacho, N.I., McIntyre, P.J., Kliebenstein, D.J., Strauss, S.Y., 2021. Genome size evolution is associated with climate seasonality and glucosinolates, but not life history, soil nutrients or range size, across a clade of mustards. Annals of Botany 127, 887–902. https://doi.org/10.1093/aob/mcab028

Cao, Z., Deng, Z., McLaughlin, M., 2014. Interspecific genome size and chromosome number variation shed new light on species classification and evolution in caladium. Journal of the American Society for Horticultural Science 139, 449–459.

Charlesworth, B., Charlesworth, D., 1983. The population dynamics of transposable elements. Genet. Res. 42, 1–27. https://doi.org/10.1017/S0016672300021455

Chen, L., Sun, J., Su, D., Cao, Q., Li, Z., Han, yonghua, 2020. Phylogenetic study of Ipomoeeae (Convolvulaceae) based on chromosome painting (preprint). In Review. https://doi.org/10.21203/rs.2.22837/v1

Cheng, Z., Dong, F., Langdon, T., Ouyang, S., Buell, C.R., Gu, M., Blattner, F.R., Jiang, J., 2002. Functional Rice Centromeres Are Marked by a Satellite Repeat and a Centromere-Specific Retrotransposon. The Plant Cell 14, 1691–1704. https://doi.org/10.1105/tpc.003079

Costa, L., Marques, A., Buddenhagen, C., Thomas, W.W., Huettel, B., Schubert, V., Dodsworth, S., Houben, A., Souza, G., Pedrosa-Harand, A., 2021. Aiming off the target: recycling target capture sequencing reads for investigating repetitive DNA. Annals of Botany mcab063. https://doi.org/10.1093/aob/mcab063

Díez, C.M., Gaut, B.S., Meca, E., Scheinvar, E., Montes-Hernandez, S., Eguiarte, L.E., Tenaillon, M.I., 2013. Genome size variation in wild and cultivated maize along altitudinal gradients. New Phytol. 199, 264–276. https://doi.org/10.1111/nph.12247

Diniz-Filho, J.A.F., de Sant'Ana, C.E.R., Bini, L.M., 1998. AN EIGENVECTOR METHOD FOR ESTIMATING PHYLOGENETIC INERTIA. Evolution 52, 1247–1262. https://doi.org/10.1111/j.1558-5646.1998.tb02006.x

Dodsworth, S., 2015. Genome skimming for next-generation biodiversity analysis. Trends in Plant Science 20, 525–527. https://doi.org/10.1016/j.tplants.2015.06.012

Dolgin, E.S., Charlesworth, B., 2008. The Effects of Recombination Rate on the Distribution and Abundance of Transposable Elements. Genetics 178, 2169–2177. https://doi.org/10.1534/genetics.107.082743

Drummond, A.J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evolutionary Biology 7, 214. https://doi.org/10.1186/1471-2148-7-214

Elliott, T.A., Gregory, T.R., 2015. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. Philosophical Transactions of the Royal Society B: Biological Sciences 370, 20140331. https://doi.org/10.1098/rstb.2014.0331

Enke, N., Fuchs, J., Gemeinholzer, B., 2011. Shrinking genomes? Evidence from genome size variation in Crepis (Compositae). Plant Biology 13, 185–193. https://doi.org/10.1111/j.1438-8677.2010.00341.x

Faizullah, L., Morton, J.A., Hersch-Green, E.I., Walczyk, A.M., Leitch, A.R., Leitch, I.J., 2021. Exploring environmental selection on genome size in angiosperms. Trends in Plant Science 26, 1039–1049. https://doi.org/10.1016/j.tplants.2021.06.001

Gale, S. (1944). "*Rhynchospora*, section *Eurhynchospora*, in Canada, the United States and the West Indies," in Contributions from the Gray Herbarium of Harvard University, 89–ii

Galindo da Costa, A.C., Thomas, W.W., Campos D. Maia, A., Navarro, D.M. do A.F., Milet-Pinheiro, P., Machado, I.C., 2021. A Continuum of Conspicuousness, Floral Signals, and Pollination Systems in Rhynchospora (Cyperaceae): Evidence of Ambophily and Entomophily in a Mostly Anemophilous Family. Ann. Mo. Bot. Gard. 106, 372–391. https://doi.org/10.3417/2021674

Galindo-González, L., Mhiri, C., Deyholos, M.K., Grandbastien, M.-A., 2017. LTR-retrotransposons in plants: Engines of evolution. Gene 626, 14–25. https://doi.org/10.1016/j.gene.2017.04.051

Gemmell, N.J., 2021. Repetitive DNA: genomic dark matter matters. Nat Rev Genet 22, 342–342. https://doi.org/10.1038/s41576-021-00354-8

Gernhard, T., 2008. New Analytic Results for Speciation Times in Neutral Models. Bull. Math. Biol. 70, 1082–1097. https://doi.org/10.1007/s11538-007-9291-0

Grandbastien, M.-A., Audeon, C., Bonnivard, E., Casacuberta, J.M., Chalhoub, B., Costa, A.-P.P., Le, Q.H., Melayah, D., Petit, M., Poncet, C., Tam, S.M., van Sluys, M.-A., Mhiri, C., 2005. Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae. Cytogenet Genome Res 110, 229–241. https://doi.org/10.1159/000084957

Grotkopp, E., Rejmánek, M., Sanderson, M.J., Rost, T.L., 2004. EVOLUTION OF GENOME SIZE IN PINES (PINUS) AND ITS LIFE-HISTORY CORRELATES: SUPERTREE

ANALYSES. Evolution 58, 1705–1729. https://doi.org/10.1111/j.0014-3820.2004.tb00456.x

Guénard, G., Legendre, P., Peres-Neto, P., 2013. Phylogenetic eigenvector maps: a framework to model and predict species traits. Methods Ecol Evol 4, 1120–1131. https://doi.org/10.1111/2041-210X.12111

Hartley, G., O'Neill, R., 2019. Centromere Repeats: Hidden Gems of the Genome. Genes 10, 223. https://doi.org/10.3390/genes10030223

Hidalgo, O., Pellicer, J., Christenhusz, M., Schneider, H., Leitch, A.R., Leitch, I.J., 2017. Is There an Upper Limit to Genome Size? Trends in Plant Science 22, 567–573. https://doi.org/10.1016/j.tplants.2017.04.005

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25, 1965–1978. https://doi.org/10.1002/joc.1276

Hijmans, R.J., van Etten, J., 2012. raster: Geographic analysis and modeling with raster data. R package version 2.0-12. http://CRAN.R-project.org/package=raster

Houben, A., Schroeder-Reiter, E., Nagaki, K., Nasuda, S., Wanner, G., Murata, M., Endo, T.R., 2007. CENH3 interacts with the centromeric retrotransposon cereba and GC-rich satellites and locates to centromeric substructures in barley. Chromosoma 116, 275–283. https://doi.org/10.1007/s00412-007-0102-z

Hu, T.T., Pattyn, P., Bakker, E.G., Cao, J., Cheng, J.-F., Clark, R.M., Fahlgren, N., Fawcett, J.A., Grimwood, J., Gundlach, H., Haberer, G., Hollister, J.D., Ossowski, S., Ottilar, R.P., Salamov, A.A., Schneeberger, K., Spannagl, M., Wang, X., Yang, L., Nasrallah, M.E., Bergelson, J., Carrington, J.C., Gaut, B.S., Schmutz, J., Mayer, K.F.X., Van de Peer, Y., Grigoriev, I.V., Nordborg, M., Weigel, D., Guo, Y.-L., 2011. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat Genet 43, 476–481. https://doi.org/10.1038/ng.807

Ibarra-Laclette, E., Lyons, E., Hernández-Guzmán, G., Pérez-Torres, C.A., Carretero-Paulet, L., Chang, T.-H., Lan, T., Welch, A.J., Juárez, M.J.A., Simpson, J., Fernández-Cortés, A., Arteaga-Vázquez, M., Góngora-Castillo, E., Acevedo-Hernández, G., Schuster, S.C., Himmelbauer, H., Minoche, A.E., Xu, S., Lynch, M., Oropeza-Aburto, A., Cervantes-Pérez, S.A., de Jesús Ortega-Estrada, M., Cervantes-Luevano, J.I., Michael, T.P., Mockler, T., Bryant, D., Herrera-Estrella, A., Albert, V.A., Herrera-Estrella, L., 2013. Architecture and evolution of a minute plant genome. Nature 498, 94–98. https://doi.org/10.1038/nature12132

Kalendar, R., Tanskanen, J., Immonen, S., Nevo, E., Schulman, A.H., 2000. Genome evolution of wild barley (Hordeum spontaneum) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proceedings of the National Academy of Sciences 97, 6603–6607. https://doi.org/10.1073/pnas.110587497

Kapitonov, V.V., Jurka, J., 2008. A universal classification of eukaryotic transposable elements implemented in Repbase. Nature Reviews Genetics 9, 411–412. https://doi.org/10.1038/nrg2165-c1

Katoh, K., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30, 3059–3066. https://doi.org/10.1093/nar/gkf436

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A., 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

Kelley, J.L., Peyton, J.T., Fiston-Lavier, A.-S., Teets, N.M., Yee, M.-C., Johnston, J.S., Bustamante, C.D., Lee, R.E., Denlinger, D.L., 2014. Compact genome of the Antarctic midge is likely an adaptation to an extreme environment. Nature Communications 5. https://doi.org/10.1038/ncomms5611

Kiazim, L.G., O'Connor, R.E., Larkin, D.M., Romanov, M.N., Narushin, V.G., Brazhnik, E.A., Griffin, D.K., 2021. Comparative Mapping of the Macrochromosomes of Eight Avian Species Provides Further Insight into Their Phylogenetic Relationships and Avian Karyotype Evolution. Cells 10, 362. https://doi.org/10.3390/cells10020362

Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16, 111–120. https://doi.org/10.1007/BF01731581

Kimura, Y., Tosa, Y., Shimada, S., Sogo, R., Kusaba, M., Sunaga, T., Betsuyaku, S., Eto, Y., Nakayashiki, H., Mayama, S., 2001. OARE-1, a Ty1-copia Retrotransposon in Oat Activated by Abiotic and Biotic Stresses. Plant and Cell Physiology 42, 1345–1354. https://doi.org/10.1093/pcp/pce171

Kraitshtein, Z., Yaakov, B., Khasdan, V., Kashkush, K., 2010. Genetic and Epigenetic Dynamics of a Retrotransposon After Allopolyploidization of Wheat. Genetics 186, 801–812. https://doi.org/10.1534/genetics.110.120790

Kükenthal, G., 1939. Vorarbeiten zu einer Monographie der Rhynchosporoideae. Feddes Repert 47, 209–216. https://doi.org/10.1002/fedr.19390471502

Liu, Z.L., Han, F.P., Tan, M., Shan, X.H., Dong, Y.Z., Wang, X.Z., Fedak, G., Hao, S., Liu, B., 2004. Activation of a rice endogenous retrotransposon Tos17 in tissue culture is accompanied by cytosine demethylation and causes heritable alteration in methylation pattern of flanking genomic regions. Theor Appl Genet 109, 200–209. https://doi.org/10.1007/s00122-004-1618-8

Lower, S.S., McGurk, M.P., Clark, A.G., Barbash, D.A., 2018. Satellite DNA evolution: old ideas, new approaches. Current Opinion in Genetics & Development 49, 70–78. https://doi.org/10.1016/j.gde.2018.03.003

Luceño, M., Vanzela, A.L., Guerra, M., 1998. Cytotaxonomic studies in Brazilian *Rhynchospora* (Cyperaceae), a genus exhibiting holocentric chromosomes. Canadian Journal of Botany 76, 440–449. https://doi.org/10.1139/b98-013

Lyu, H., He, Z., Wu, C.-I., Shi, S., 2018. Convergent adaptive evolution in marginal environments: unloading transposable elements as a common strategy among mangrove genomes. New Phytologist 217, 428–438. https://doi.org/10.1111/nph.14784

Macas, J., Novák, P., Pellicer, J., Čížková, J., Koblížková, A., Neumann, P., Fuková, I., Doležel, J., Kelly, L.J., Leitch, I.J., 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, e0143424. https://doi.org/10.1371/journal.pone.0143424

Madlung, A., Tyagi, A.P., Watson, B., Jiang, H., Kagochi, T., Doerge, R.W., Martienssen, R., Comai, L., 2004. Genomic changes in synthetic Arabidopsis polyploids: Genomic changes in Arabidopsis polyploids. The Plant Journal 41, 221–230. https://doi.org/10.1111/j.1365-313X.2004.02297.x

Marques, A., Ribeiro, T., Neumann, P., Macas, J., Novák, P., Schubert, V., Pellino, M., Fuchs, J., Ma, W., Kuhlmann, M., Brandt, R., Vanzela, A.L.L., Beseda, T., Šimková, H., Pedrosa-Harand, A., Houben, A., 2015. Holocentromeres in *Rhynchospora* are associated with genome-wide centromere-specific repeat arrays interspersed among euchromatin. Proceedings of the National Academy of Sciences 112, 13633–13638. https://doi.org/10.1073/pnas.1512255112

Márquez-Corro, J.I., Martín-Bravo, S., Pedrosa-Harand, A., Hipp, A.L., Luceño, M., Escudero, M., 2019. Karyotype Evolution in Holocentric Organisms, in: John Wiley & Sons, Ltd (Ed.), ELS. Wiley, pp. 1–7. https://doi.org/10.1002/9780470015902.a0028758

Mascagni, F., Vangelisti, A., Giordani, T., Cavallini, A., Natali, L., 2020. A computational comparative study of the repetitive DNA in the genus Quercus L. Tree Genetics & Genomes 16, 11. https://doi.org/10.1007/s11295-019-1401-2

Mata-Sucre, Y., Sader, M., Van-Lume, B., Gagnon, E., Pedrosa-Harand, A., Leitch, I.J., Lewis, G.P., Souza, G., 2020. How diverse is heterochromatin in the Caesalpinia group? Cytogenomic characterization of Erythrostemon hughesii Gagnon & G.P. Lewis (Leguminosae: Caesalpinioideae). Planta 252, 49. https://doi.org/10.1007/s00425-020-03453-8

Matsunaga, W., Ohama, N., Tanabe, N., Masuta, Y., Masuda, S., Mitani, N., Yamaguchi-Shinozaki, K., Ma, J.F., Kato, A., Ito, H., 2015. A small RNA mediated regulation of a stress-activated retrotransposon and the tissue specific transposition during the reproductive period in Arabidopsis. Front. Plant Sci. 6. https://doi.org/10.3389/fpls.2015.00048

McClintock, B., 1984. The Significance of Responses of the Genome to Challenge. Science 226, 792–801. https://doi.org/10.1126/science.15739260

Mehrotra, S., Goyal, V., 2014. Repetitive Sequences in Plant Nuclear DNA: Types, Distribution, Evolution and Function. Genomics, Proteomics & Bioinformatics 12, 164–171. https://doi.org/10.1016/j.gpb.2014.07.003

Meštrović, N., Mravinac, B., Pavlek, M., Vojvoda-Zeljko, T., Šatović, E., Plohl, M., 2015. Structural and functional liaisons between transposable elements and satellite DNAs. Chromosome Res 23, 583–596. https://doi.org/10.1007/s10577-015-9483-7

Navajas-Pérez, R., Schwarzacher, T., Ruiz Rejón, M., Garrido-Ramos, M.A., 2009. Characterization of RUSI, a telomere-associated satellite DNA, in the genus *Rumex* (Polygonaceae). Cytogenet Genome Res 124, 81–89. https://doi.org/10.1159/000200091

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. https://doi.org/10.1186/s13100-018-0144-1

Nimon, K., Oswald, F., Roberts, J.K., 2021. yhat: Interpreting Regression Effects. R package version 2.0-3. https://CRAN.R-project.org/package=yhat

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 Research 45, e111–e111. https://doi.org/10.1093/nar/gkx257

Novák, P., Guignard, M.S., Neumann, P., Kelly, L.J., Mlinarec, J., Koblížková, A., Dodsworth, S., Kovařík, A., Pellicer, J., Wang, W., Macas, J., Leitch, I.J., Leitch, A.R., 2020. Repeat-sequence turnover shifts fundamentally in species with large genomes. Nat. Plants 6, 1325–1329. https://doi.org/10.1038/s41477-020-00785-x

Novak, 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, 792–793. https://doi.org/10.1093/bioinformatics/btt054

Oliveira da Silva, W., Malcher, S.M., Pereira, A.L., Pieczarka, J.C., Ferguson-Smith, M.A., O'Brien, P.C.M., Mendes-Oliveira, A.C., Geise, L., Nagamachi, C.Y., 2020. Chromosomal Signatures Corroborate the Phylogenetic Relationships within Akodontini (Rodentia, Sigmodontinae). IJMS 21, 2415. https://doi.org/10.3390/ijms21072415

Osborne, C.P., Sack, L., 2012. Evolution of C 4 plants: a new hypothesis for an interaction of CO 2 and water relations mediated by plant hydraulics. Philos. Trans. R. Soc. B Biol. Sci. 367, 583–600. https://doi.org/10.1098/rstb.2011.0261

Pandit, M.K., White, S.M., Pocock, M.J.O., 2014. The contrasting effects of genome size, chromosome number and ploidy level on plant invasiveness: a global analysis. New Phytol 203, 697–703. https://doi.org/10.1111/nph.12799

Parisod, C., Alix, K., Just, J., Petit, M., Sarilar, V., Mhiri, C., Ainouche, M., Chalhoub, B., Grandbastien, M.-A., 2010. Impact of transposable elements on the organization and function of allopolyploid genomes: Research review. New Phytologist 186, 37–45. https://doi.org/10.1111/j.1469-8137.2009.03096.x

Petit, M., Guidat, C., Daniel, J., Denis, E., Montoriol, E., Bui, Q.T., Lim, K.Y., Kovarik, A., Leitch, A.R., Grandbastien, M.-A., Mhiri, C., 2010. Mobilization of retrotransposons in synthetic allotetraploid tobacco. New Phytologist 186, 135–147. https://doi.org/10.1111/j.1469-8137.2009.03140.x

Petrov, D.A., Fiston-Lavier, A.-S., Lipatov, M., Lenkov, K., Gonzalez, J., 2011. Population Genomics of Transposable Elements in Drosophila melanogaster. Molecular Biology and Evolution 28, 1633–1644. https://doi.org/10.1093/molbev/msq337

Plohl, M., Meštrović, N., Mravinac, B., 2014. Centromere identity from the DNA point of view. Chromosoma 123, 313–325. https://doi.org/10.1007/s00412-014-0462-0

Plummer, M., Best, N., Cowles, K., Vines, K., 2006. CODA: convergence diagnosis and output analysis for MCMC. R News 6, 7–11.

R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rabosky, D.L., 2014. Automatic Detection of Key Innovations, Rate Shifts, and Diversity-Dependence on Phylogenetic Trees. PLoS ONE 9, e89543. https://doi.org/10.1371/journal.pone.0089543

Ramallo, E., Kalendar, R., Schulman, A.H., Martínez-Izquierdo, J.A., 2008. Reme1, a Copia retrotransposon in melon, is transcriptionally induced by UV light. Plant Mol Biol 66, 137–150. https://doi.org/10.1007/s11103-007-9258-4

Rambaut, A., Drummond, A.J., 2013. TreeAnnotator v1. 7.0. Available as Part of the BEAST package.

Ray-Mukherjee, J., Nimon, K., Mukherjee, S., Morris, D.W., Slotow, R., Hamer, M., 2014. Using commonality analysis in multiple regressions: a tool to decompose regression effects in the face of multicollinearity. Methods Ecol Evol 5, 320–328. https://doi.org/10.1111/2041-210X.12166

Renny-Byfield, S., Kovarik, A., Kelly, L.J., Macas, J., Novak, P., Chase, M.W., Nichols, R.A., Pancholi, M.R., Grandbastien, M.-A., Leitch, A.R., 2013. Diploidization and genome size change in allopolyploids is associated with differential dynamics of low- and high-copy sequences. Plant J 74, 829–839. https://doi.org/10.1111/tpj.12168

Revell, L.J., 2012. phytools: an R package for phylogenetic comparative biology (and other things): *phytools: R package*. Methods in Ecology and Evolution 3, 217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x

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

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

Ribeiro, T., Vasconcelos, E., dos Santos, K.G.B., Vaio, M., Brasileiro-Vidal, A.C., Pedrosa-Harand, A., 2020. Diversity of repetitive sequences within compact genomes of

Phaseolus L. beans and allied genera Cajanus L. and Vigna Savi. Chromosome Res 28, 139–153. https://doi.org/10.1007/s10577-019-09618-w

Robberecht, C., Voet, T., Esteki, M.Z., Nowakowska, B.A., Vermeesch, J.R., 2013. Nonallelic homologous recombination between retrotransposable elements is a driver of de novo unbalanced translocations. Genome Research 23, 411–418. https://doi.org/10.1101/gr.145631.112

Sakamoto, M., Venditti, C., 2018. Phylogenetic non-independence in rates of trait evolution. Biol. Lett. 14, 20180502. https://doi.org/10.1098/rsbl.2018.0502

Santos, T., 2018. PVR: Phylogenetic Eigenvectors Regression and Phylogentic Signal-Representation Curve. R package version 0.3. https://CRAN.R-project.org/package=PVR

Schley, R.J., Pellicer, J., Ge, X.-J., Barrett, C., Bellot, S., S. Guignard, M., Novák, P., Suda, J., Fraser, D., Baker, W.J., Dodsworth, S., Macas, J., Leitch, A.R., Leitch, I.J., 2021. The Ecology of Palm Genomes: Repeat-associated genome size expansion is constrained by aridity (preprint). Evolutionary Biology. https://doi.org/10.1101/2021.11.04.467295

Silva Filho, P.J.S., Thomas, W.W., Boldrini, I.I., 2021. Redefining *Rhynchospora* section *Tenues* (Cyperaceae), a phylogenetic approach. Botanical Journal of the Linnean Society boab002. https://doi.org/10.1093/botlinnean/boab002

Souza, G., Costa, L., Guignard, M.S., Van-Lume, B., Pellicer, J., Gagnon, E., Leitch, I.J., Lewis, G.P., 2019. 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 38, 13–23. https://doi.org/10.1016/j.ppees.2019.03.002

Spalink, D., Drew, B.T., Pace, M.C., Zaborsky, J.G., Starr, J.R., Cameron, K.M., Givnish, T.J., Sytsma, K.J., 2016. Biogeography of the cosmopolitan sedges (Cyperaceae) and the area-richness correlation in plants. Journal of Biogeography 43, 1893–1904. https://doi.org/10.1111/jbi.12802

Thomas, W.W., 2020. Two new species of Rhynchospora (Cyperaceae) from Bahia, Brazil, and new combinations in Rhynchospora section Pleurostachys. Brittonia 72, 273–281. https://doi.org/10.1007/s12228-020-09621-0

Trávníček, P., Čertner, M., Ponert, J., Chumová, Z., Jersáková, J., Suda, J., 2019. Diversity in genome size and GC content shows adaptive potential in orchids and is closely linked to partial endoreplication, plant life-history traits and climatic conditions. New Phytol 224, 1642–1656. https://doi.org/10.1111/nph.15996

Usai, G., Mascagni, F., Natali, L., Giordani, T., Cavallini, A., 2017. Comparative genome-wide analysis of repetitive DNA in the genus Populus L. Tree Genetics & Genomes 13, 96. https://doi.org/10.1007/s11295-017-1181-5

Vanzela, A.L.L., Guerra, M., 2000. Heterochromatin differentiation in holocentric chromosomes of Rhynchospora (Cyperaceae). Genet. Mol. Biol. 23, 453–456. https://doi.org/10.1590/S1415-47572000000200034

Vitales, D., Garcia, S., Dodsworth, S., 2019. Reconstructing Phylogenetic Relationships Based on Repeat Sequence Similarities. bioRxiv. https://doi.org/10.1101/624064

Vojvoda Zeljko, T., Pavlek, M., Meštrović, N., Plohl, M., 2020. Satellite DNA-like repeats are dispersed throughout the genome of the Pacific oyster Crassostrea gigas carried by Helentron non-autonomous mobile elements. Sci Rep 10, 15107. https://doi.org/10.1038/s41598-020-71886-y

Yaakov, B., Kashkush, K., 2012. Mobilization of Stowaway-like MITEs in newly formed allohexaploid wheat species. Plant Mol Biol 80, 419–427. https://doi.org/10.1007/s11103-012-9957-3

Yang, X., Zhao, H., Zhang, T., Zeng, Z., Zhang, P., Zhu, B., Han, Y., Braz, G.T., Casler, M.D., Schmutz, J., Jiang, J., 2018. Amplification and adaptation of centromeric repeats in polyploid switchgrass species. New Phytol 218, 1645–1657. https://doi.org/10.1111/nph.15098

Yu, F., Dou, Q., Liu, R., Wang, H., 2017. A conserved repetitive DNA element located in the centromeres of chromosomes in Medicago genus. Genes Genom 39, 903–911. https://doi.org/10.1007/s13258-017-0556-1

Zhong, C.X., Marshall, J.B., Topp, C., Mroczek, R., Kato, A., Nagaki, K., Birchler, J.A., Jiang, J., Dawe, R.K., 2002. Centromeric Retroelements and Satellites Interact with Maize Kinetochore Protein CENH3. The Plant Cell 14, 2825–2836. https://doi.org/10.1105/tpc.006106

Zhu, Q.-H., Shan, W.-X., Ayliffe, M.A., Wang, M.-B., 2016. Epigenetic Mechanisms: An Emerging Player in Plant-Microbe Interactions. MPMI 29, 187–196. https://doi.org/10.1094/MPMI-08-15-0194-FI

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., Farooq, H., Herdean, A., Ariza, M., Scharn, R., Svantesson, S., Wengström, N., Zizka, V., Antonelli, A., 2019. CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. Methods Ecol Evol 10, 744–751. https://doi.org/10.1111/2041-210X.13152

Table 1 – Predictor variables used in this study, with *p*-values, adjusted R^2, F-statistics and Coefficient (b) of all simple linear regression models. Variables in bold were selected for the multiple linear regression and commonality analyses

Predictor	***p*-value**	**R^2**	**F-statistic**	**Coefficient (β)**
Tyba abundance	**< 0.001**	**0.31**	**30.84**	**-0.216**
Age of MRA	**< 0.001**	**0.21**	**19.55**	**0.045**
Phylogenetic Eigenvectors	**< 0.001**	**0.47**	**60.56**	**2.79**
Latitude	0.001	0.13	11.42	-0.008
Mean Diurnal Range (Bio 2)	0.06	0.04	3.62	-0.007
Temperature Annual Range (Bio 7)	< 0.001	0.19	17.3	-0.003
Precipitation of Wettest Month (Bio 13)	**< 0.001**	**0.22**	**19.86**	**0.003**
Precipitation of Driest Month (Bio 14)	0.01	0.07	6.28	-0.005
Precipitation Seasonality (Bio 15)	< 0.001	0.2	18.06	0.01
Precipitation of Warmest Quarter (Bio 18)	0.64	-0.01	0.228	-0.0002
Precipitation of Coldest Quarter (Bio 19)	0.31	0.001	1.066	-0.0003

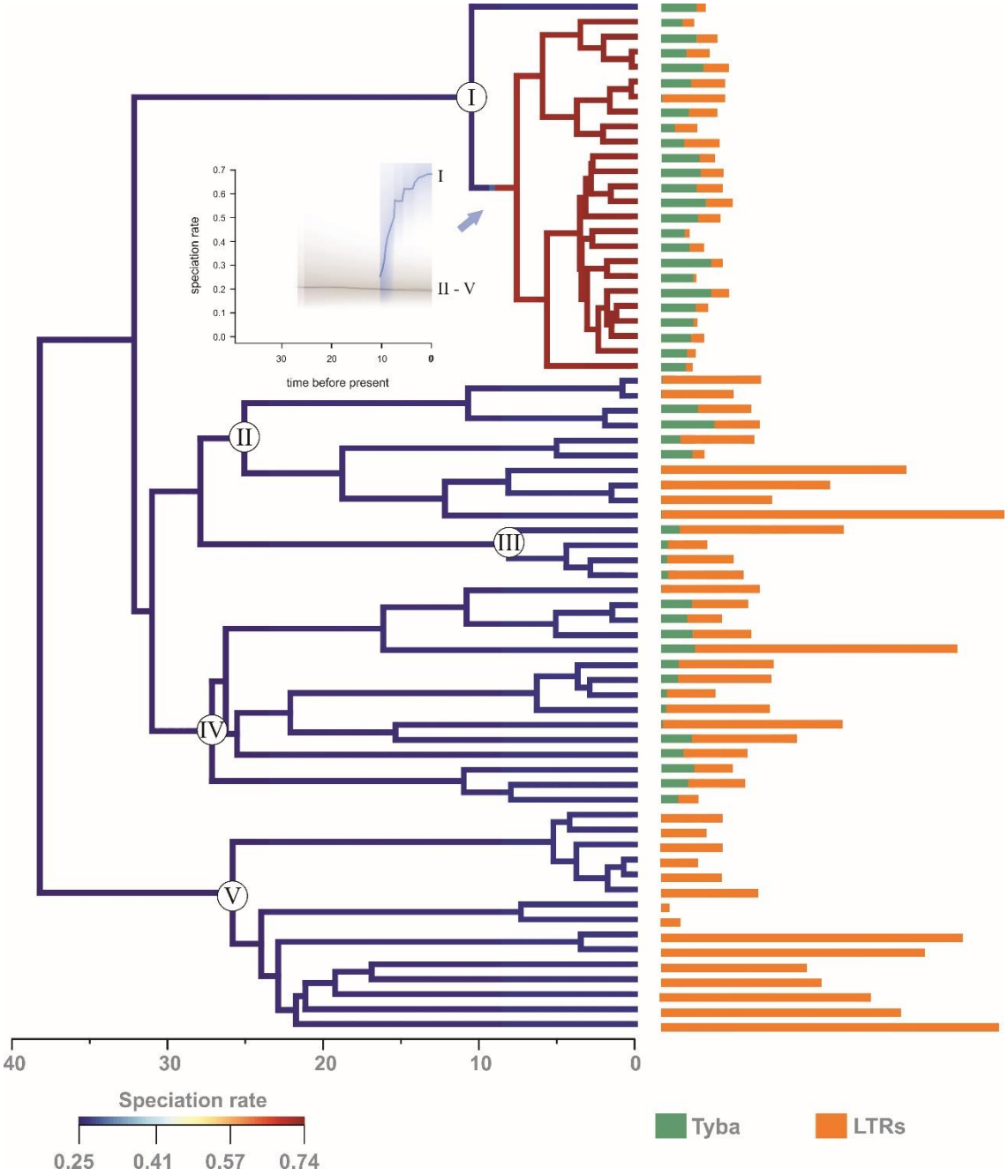


Figure 1 - Dated phylogenetic tree of 69 *Rhynchospora* species. Colors of the branches represent the speciation rate as given by BAMM following the legend at the lower left corner. Circles with roman numbers in the nodes delimits the four clades that are discussed throughout the text. Bars on the right of the tips represent the genomic abundance of *Tyba* (green) and LTR retrotransposons (orange). Blue arrow represents the node in which it was detected a shift on the diversification rate by the BAMM analysis. The speciation through time (plot directly above it) of clade I (blue outline) was much higher compared to other clades (grey outline).

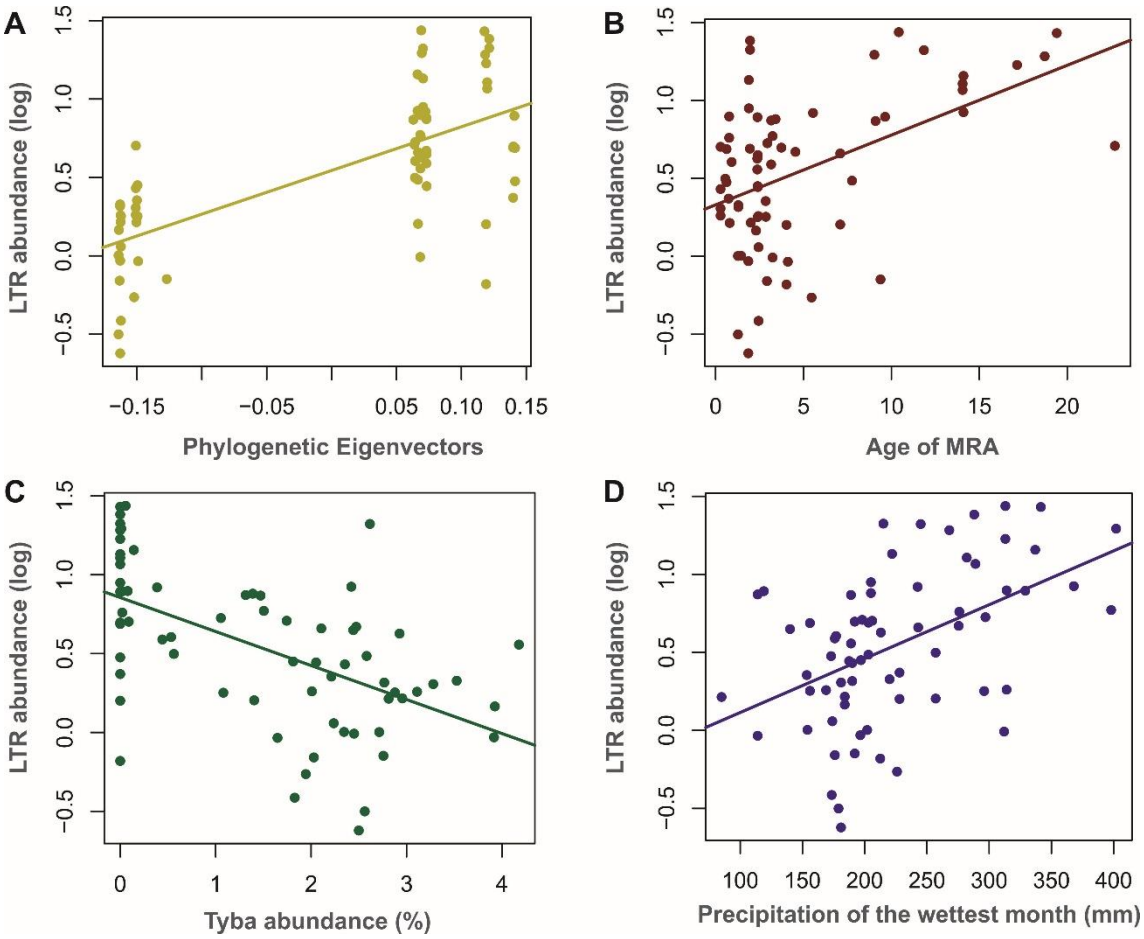


Figure 2 – Graphics with regression lines for LTR-RT abundance (%) in function of (A) Phylogenetic Eigenvectors, (B) age of most recente ancestor (MRA), (C) *Tyba* abundance (%) and (D) precipitation of the wettest month (mm).

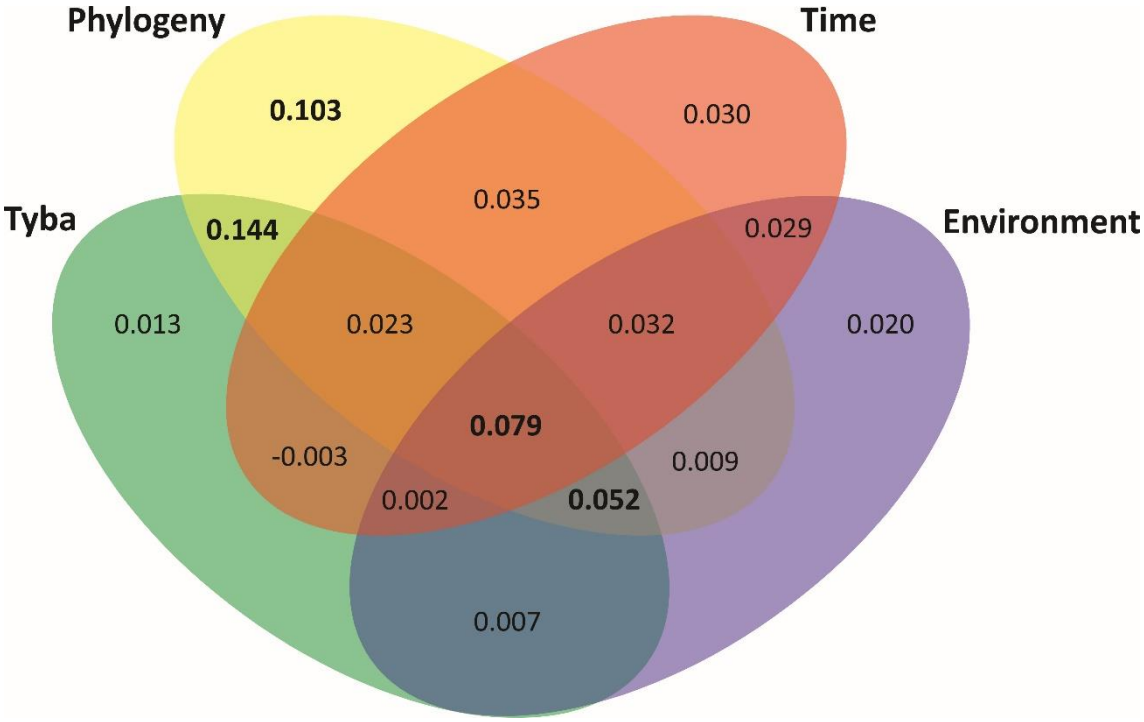


Figure 3 - Venn diagram representing the results from the commonality analysis based on the multiple regression analysis of LTR-RT abundance in function of *Tyba* abundance, phylogeny (phylogenetic eigenvectors), time (age of most recente ancestor) and environment (precipitation of wettest month). Values within ellipses represent the adjusted R^2 values of the individual and shared contribution of the variables to the strength of the regression model. Contributions above 5% are highlighted in bold.

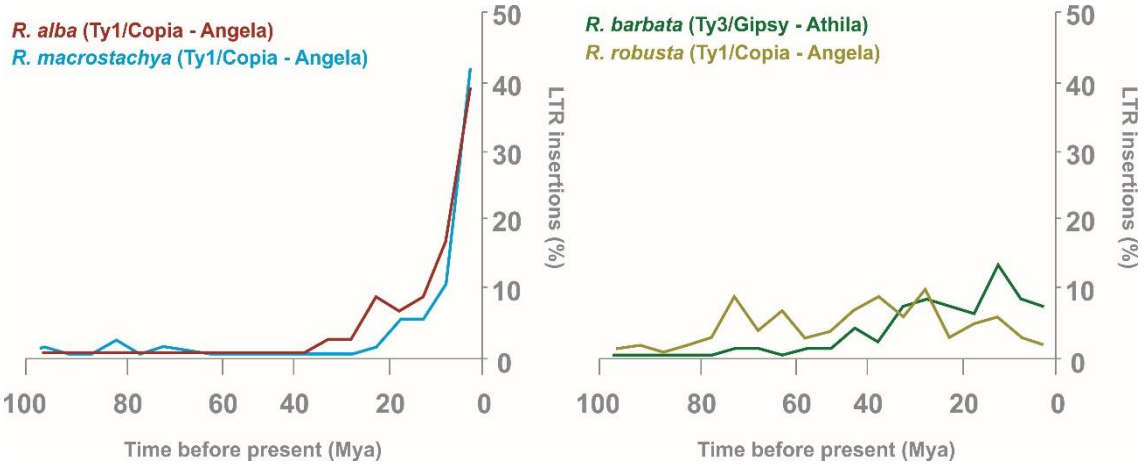


Figure 4 – Frequency histogram representing the proportion of insertions along the years for the most abundant element for species with low LTR-RT abundance [*R. alba* (red) and *R. macrostachya* (blue)] on the left and species with high LTR-RT abundance [*R. barbata* (green), and *R. robusta* (yellow)] on the right.

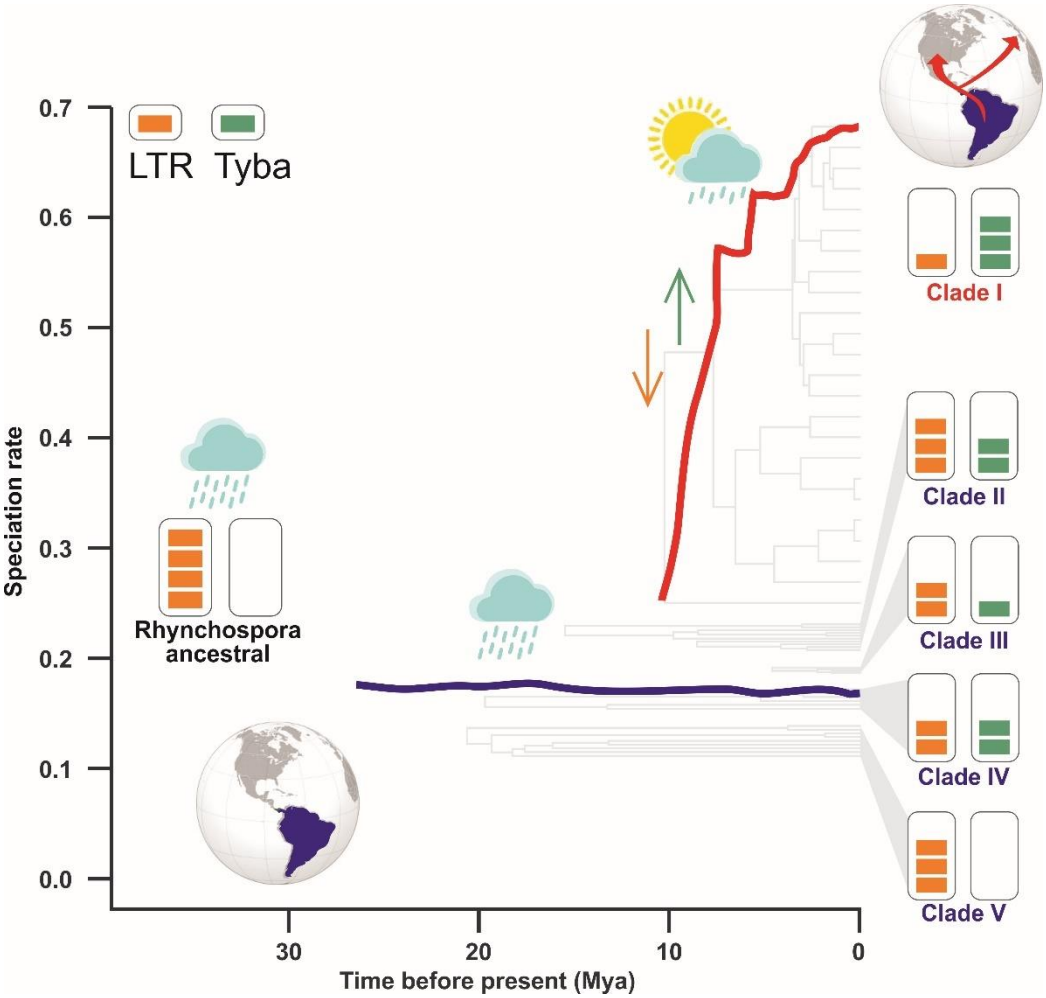


Figure 5 –LTR-RT evolution in *Rhynchopora*. The genus ancestor originated in rainier areas of South America, with posterior dispersion to low-precipitation areas of the Northern Hemisphere (lower left and upper right globes, respectively). This migration, reflected by the phylogenetic relationships (light grey) was accompanied by a boom of speciation rate (y-axis), loss of LTR-RTs (orange) and amplification of *Tyba* sequences (green).

Supplementary Table 1 – List of *Rhynchospora* species, voucher number, herbarium, total number of reads analyzed in the RepeatExplorer analysis, LTR-RT and *Tyba* proportion on the genome (%).

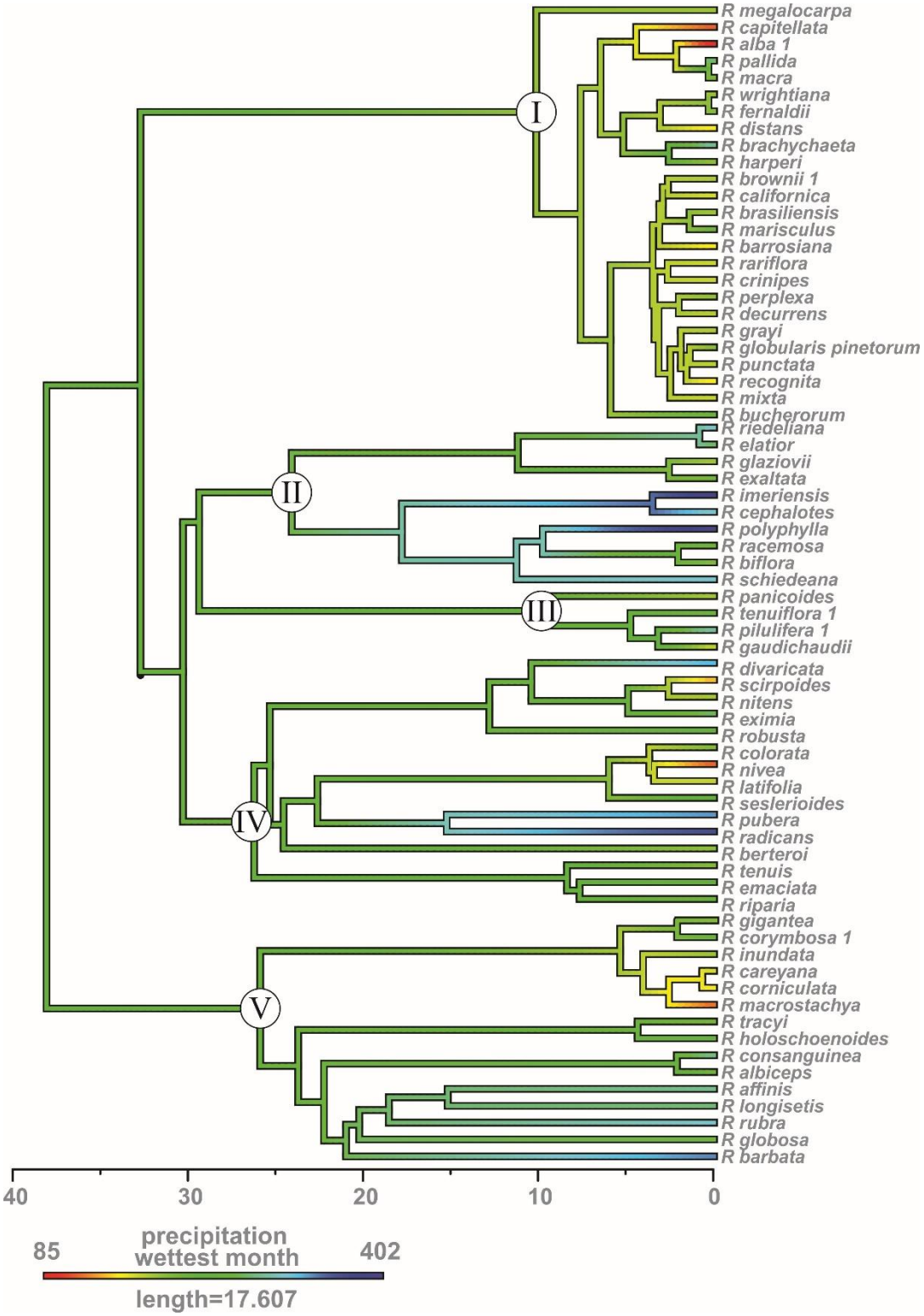
Species	Voucher – herbarium	Analyzed reads	LTR-RT proportion (%)	*Tyba* proportion
Carex annectens (E.P.Bicknell) P.Bicknell	Andrew Hipp 2013014 – MOR	-	-	-
Carex lupulina Muhl. ex Willd.	Andrew Hipp 2013012 – MOR	-	-	-
Carex lúrida Wahlenb.	Loran C. Anderson 24918 – FSU	-	-	-
Carex scoparia Willd.	Marilee Lovit 464 – MOR	-	-	-
Carex tribuloides Wahlenb.	Andrew Hipp 2013001 – MOR	-	-	-
Carex waponahkikensis M.Lovit & A.Haines	Marilee Lovit 420B - MOR	-	-	-
Chorizandra multiarticulata Nees	Chrissie J. Prychid 40 – NE	-	-	-
Chorizandra sphaerocephala R.Br.	Chrissie J. Prychid 42 – NE	-	-	-
Exocarya sclerioides (F. Muell.) Benth	Chrissie J. Prychid 41 – NE	-	-	-
Hypolytrum nemorum (Vahl)	Chrissie J. Prychid 45 – NE	-	-	-
R. affinis W.Fitzg.	A.W. Scott 79225 – NE	1,516,574	11.40	0
R. alba (L.) Vahl	Bob Moyer – FSU	747,026	1.60	2.81
R. albiceps Kunth	Wayt Thomas 16397 – NY	810,371	20.65	0
R. barbata (Vahl) Kunth	Wayt Thomas 16195 – NY	417,260	26.41	0
R. barrosiana Guagl.	Jeremy Bruhl 2330 – NE	762,091	1.75	2.88
R. berteroi (Spreng.) C.B.Clarke	Wayt Thomas 14864 - NY	978,959	4.99	1.74
R. biflora Boeckeler	Wayt Thomas 15002 – NY	1,104,253	8.70	0
R. brachychaeta Wright ex Sauv.	Orzell and Bridges 26718 – FSU	814,765	1.75	1.08
R. brasiliensis Boeckeler	Wayt Thomas 16402 – NY	623,699	2.02	2.77
R. brownii Roem. & Schult.	L.M. Copeland 3336 – NE	826,435	1.61	2.96
R. bucherorum León	Wayt Thomas 14933 – NY	1131705	0.53	1.94
R. californica Gale	Robert Naczi 10626 – DOV	621,597	1.77	3.11
R. capitellata (Michx.) Vahl	Bob Moyer s.n. – FSU	794,488	0.90	1.65
R. careyana (Michx.) Vahl	Chris Buddenhagen 1309072 – FSU	465,498	2.92	0
R. cephalotes (L.). Vahl	Robert Naczi 11691 – NY	580,881	0.96	2.45
R. colorata (L.) H. Pfeiffer	Chris Buddenhagen 1307091 – FSU	866,372	7.42	1.39
R. consanguinea (Kunth) Boeckeler	Wayt Thomas 16404 – NY	653,574	23.62	0
R. corniculata (Lam.) A.Gray	Loran C. Anderson 27381 - FSU	688,746	4.76	0
R. corymbosa (L.) Britton	Jeremy Bruhl 2316 – NE	221,871	2.29	0
R. crinipes Gale	Chris Buddenhagen 1407101 – FSU	751,962	1.12	2.24
R. decurrens Chapman	Loran Anderson 23273 – FSU	688,873	0.23	2.50
R. distans (Michaux) Vahl	Chris Buddenhagen 14082120 – FSU	529,996	2.21	2.21

R. divaricata (Ham.) M.T.Strong	Robert Naczi 12107 - NY	858,680	7.67	0.02
R. elatior Kunth	Wayt Thomas 16400 – NY	877,377	5.63	0.02
R. emaciata (Nees) Boeckeler	Jeremy Bruhl 2320 – NE	556,786	4.46	2.11
R. exaltata Kunth	Wayt Thomas s.n – NY	416,693	4.14	2.93
R. eximia (Nees von Esenbeck) Boeckeler	Chris Buddenhagen 14082224 – FSU	811,811	4.57	2.47
R. fernaldii Gale	Richard Carter 21415 – VSC	646,765	4.93	0.09
R. gaudichaudii (Brongn.) L.B. Sm.	Wayt Thomas 16319 - NY	713,956	3.94	0.53
R. gigantea Link	Wayt Thomas 16407 – NY	384,450	4.79	0
R. glaziovii Boeckeler	M Reginato 1486 – NY	173,544	3.52	4.18
R. globosa (Kunth) Roem. & Schult.	Jeremy Bruhl 2319 - NE	199,121	18.76	0
R. globularis (Chapm.) Small var. *pinetorum* (Britton & Small) Gale	Chris Buddenhagen 1107084 - FSU	592,819	0.98	2.71
R. grayi Kunth	Chris Buddenhagen 1305106 – FSU	482,515	1.43	3.93
R. harperi Small	Chris Buddenhagen 14082233 – FSU	696,277	2.76	1.81
R. holoschoenoides (Rich.) Herter	Wayt Thomas 16306B – NY	795,612	1.55	0
R. imeriensis (Kük.) W.W.Thomas	Julian Aguirre 1899 – NY	971,455	5.77	1.51
R. inundata (Oakes) Fernald	Chris Buddenhagen 13101032 – FSU	443,639	4.86	0
R. latifolia (Baldwin ex Elliott) W.W.Thomas	Loran C. Anderson 27371 – FSU	881,513	3.80	0.44
R. longisetis R.Br.	K.L. Wilson 9684 – NSW	280,947	12.51	0
R. macra (C. B. Clarke ex Britton) Small	Wayt Thomas 14709 – NY	980,680	1.98	3.28
R. macrostachya Torr. ex A.Gray	Robert Naczi 12032 – NY	676,729	7.63	0
R. marisculus Lindl. & Nees	Jeremy Bruhl 2328 – NE	773,054	2.08	3.53
R. megalocarpa A.Gray	Chris Buddenhagen 1107085 – FSU	884,916	0.69	2.76
R. mixta Britton ex Small	Loran Anderson 25574 – FSU	515,468	0.68	2.03
R. nitens (Vahl) A.Gray	Loran C. Anderson 25687 – FSU	339,281	2.72	2.05
R. nivea Boeckeler	Taylor et al 2932 – BRIT	1,020,456	7.27	1.31
R. pallida M.A.Curtis	Wayt Thomas 14707 – NY	872,903	1.78	2.01
R. panicoides Schrad. ex Nees	ER Guaglione & Sobral 19990724 – NY	300,769	7.21	1.47
R. perplexa Britto n ex Small	Loran Anderson 25485 – FSU	1,153,277	0.91	3.92
R. pilulifera Bertol.	M Reginato 1476 – NY	462,259	5.20	1.05
R. polyphylla (Vahl) Vahl	Wayt Thomas 16483 – NY	1,251,776	19.17	0.01
R. pubera (Vahl) Boeckeler	ACG Costa sn – NY	2,797,919	14.04	0.14
R. punctata Elliott	Richard Carter 21712 – VSC	634,965	0.31	2.56
R. racemosa C.Wright ex Sauvalle	Wayt Thomas 14866 - NY	728,267	13.21	0

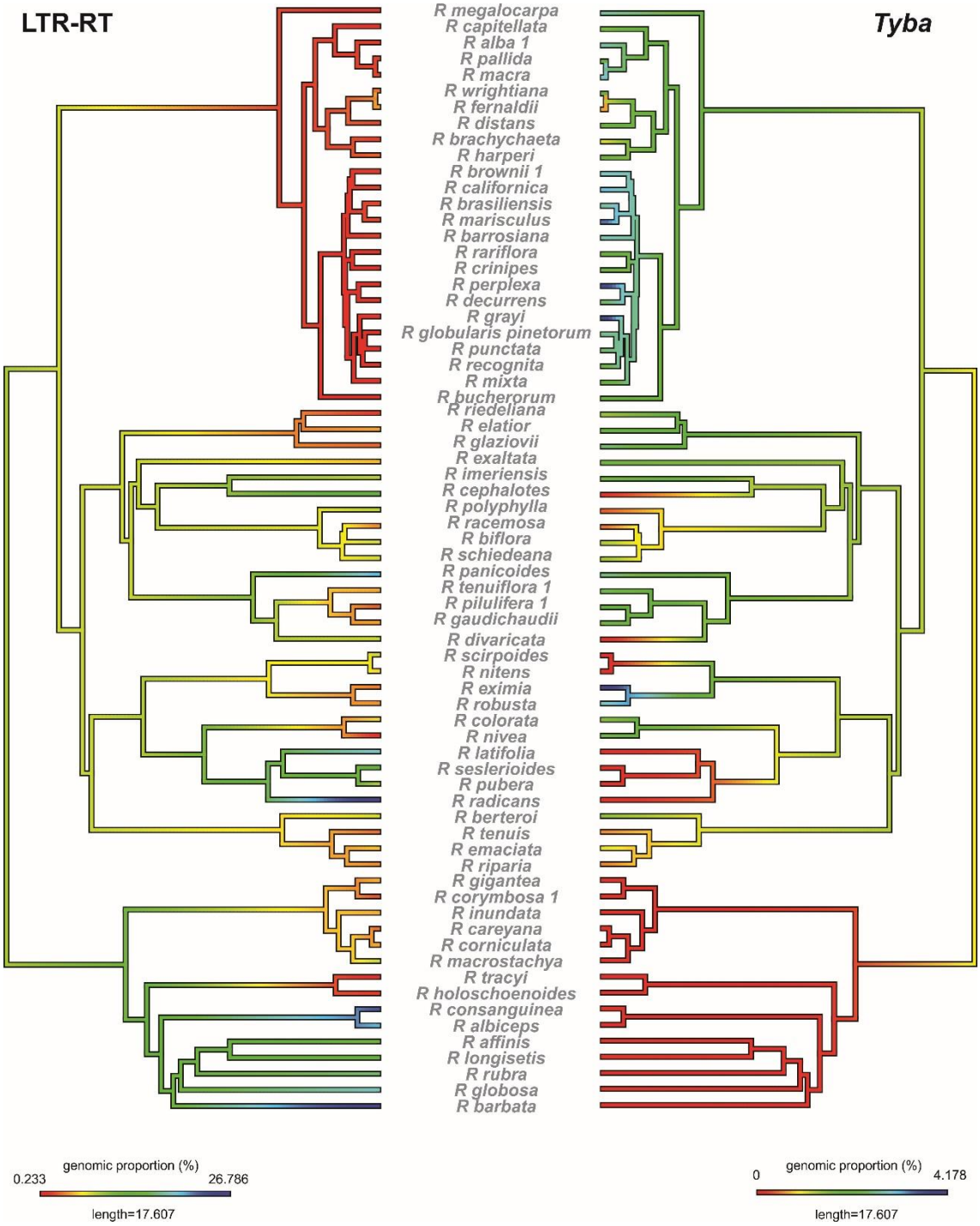
R. radicans (Schltdl. & Cham.) H.Pfeiff.	Wayt Thomas 14907 – NY	926,007	8.21	2.42
R. rariflora (Michaux) Elliott	Chris Buddenhagen 14082219 - FSU	873,893	0.38	1.83
R. recognita (Gale) Kral	Chris Buddenhagen 14082229 – FSU	582,753	0.98	2.35
R. riedeliana C.B.Clarke	Wayt Thomas 16401 - NY	1,258,905	7.71	0.08
R. riparia (Nees) Boec keler	Wayt Thomas 16153 – NY	391,118	1.56	1.40
R. robusta (Kunth) Boeckeler	Wayt Thomas 16403 – NY	650,633	20.53	2.62
R. rubra (Lour.) Makino	Jeremy Bruhl 2487 – NE	811,006	16.50	0
R. schiedeana (Schltdl.) Kunth	Wayt Thomas 16473 – NY	721,677	26.79	0.06
R. scirpoides (Torrey) A. Gray	Robert Naczi 12036 – NY	523,231	4.36	2.44
R. seslerioides Griseb.	Wayt Thomas 14924 – NY	813,824	8.13	0.39
R. tenuis Link	Wayt Thomas sn – NY	661,128	2.99	2.58
R. tenuiflora (Brongn.) L.B. Sm.	Wayt Thomas 16317 – NY	261,844	3.08	0.56
R. wrightiana Boeckeler	Richard Carter 21416 - VSC	447,408	2.64	2.35
Scirpodendron ghaeri (Gaertn.) Merr.	Chrissie J. Prychid 47 – NE	-		

Supplementary Table 2 – Estimated total number of species per each major clade based on Kükenthal (1939) section treatment and number of taxa sampled per major clade. This information was used to estimate the proportion of sampled species per clade for the BAMM analysis.

Clade	Kükenthal's sections	No. of taxa in treatment	No. of sampled taxa
I	*Albae, Cernuae, Chapmaniae, Cubenses, Fuscae, Globulares, Harveyae, Laevinuces, Marisculae, Mixtae, Plumosae, Prolifearae, Pseudo-aureae, Spermodontes, Stenophyllae, Volderugosae*	93	25
II	*Cephalotae, Paniculatae, Pseudocapitatde, Racemosae*	43	10
III	*Pleurostachys*	30	5
IV	*Dichromena, EuPsilocarya, Luzuliformes, Tenues*	52	15
V	*Longirostres, Pauciflorae, Pluriflorae*	49	15



Supplementary Figure 1 – Phylogenetic tree of *Rhynchopora* with median values of Bio13 (Precipitation of Wettest Month) plotted along the branches according to the legend at the lower left corner.



Supplementary Figure 2 – Phylogenetic trees of *Rhynchopora* with values of LTR-RT (left) and *Tyba* (right) abundances plotted along the branches according to the legends at the lower left and lower right corner respectively.

4 CONCLUSÕES

1. Foi possível caracterizar a fração repetitiva de cinco espécies de *Rhynchospora* utilizando o subproduto de sequenciamento por captura de alvo, obtendo dados comparáveis aos obtidos por *genome skimming*. Além disso, foi possível usar este subproduto para o desenvolvimento de sonda de DNA satélite para estudos citogenômicos e para a construção de árvores filogenéticas baseadas na abundância e similaridade de elementos repetitivos. Com a disponibilidade cada vez maior de sequências obtidas por captura de alvo, estes resultados mostram um amplo potencial para a reciclagem de sequências não-alvo para uma melhor caracterização da biodiversidade.

2. Foi observado que o DNA satélite holocentrômero-específico *Tyba* está presente em quatro dos cinco principais clados de *Rhynchospora* e que este apresenta pequenas variações estruturais clado/subclado específica. No geral, *Tyba* mostrou uma grande conservação de sequência, com aproximadamente 70% de similaridade entre sequências de espécies com mais de 30 milhões de anos de divergência. Padrões de curvatura das sequências de *Tyba* indicam que esta possa conferir vantagens para a formação de nucleossomos, o que pode indicar uma possível função estrutural deste DNA satélite nos holocentrômeros de *Rhynchospora*.

3. A análise de regressão entre as abundâncias de retrotransposons-LTR e *Tyba* demonstraram um padrão contrastante, sugerindo que pressões evolutivas agem de forma diferente nessas duas frações genômicas. Além disso, uma diminuição no conteúdo de LTRs foi observada em espécies de *Rhynchospora* que migraram (se diversificando rapidamente) para ambientes com menor índice de precipitação, sugerindo uma possível seleção purificadora desses elementos. Em uma análise multivariada, foi possível ver que fatores como abundância de *Tyba*, precipitação e tempo de divergência promovem um maior impacto na abundância de LTRs quando combinados com a informação filogenética, implicando uma possível relação entre a diversificação de espécies e esta fração genômica.

REFERÊNCIAS

ACHREM, M.; SZUĆKO, I.; KALINKA, A. The epigenetic regulation of centromeres and telomeres in plants and animals. **Comparative Cytogenetics**, v. 14, n. 2, p. 265–311, 7 jul. 2020.

AHMAD, S. F. et al. Dark Matter of Primate Genomes: Satellite DNA Repeats and Their Evolutionary Dynamics. **Cells**, v. 9, n. 12, p. 2714, 18 dez. 2020.

ALBERT, T. J. et al. Direct selection of human genomic loci by microarray hybridization. **Nature Methods**, v. 4, n. 11, p. 903–905, nov. 2007.

ALIX, K. et al. Polyploidy and interspecific hybridization: partners for adaptation, speciation and evolution in plants. **Annals of Botany**, v. 120, n. 2, p. 183–194, 1 ago. 2017.

ALLEN, J. M. et al. Phylogenomics from Whole Genome Sequences Using aTRAM. **Systematic Biology**, p. syw105, 25 jan. 2017.

ALROY, J. The Shifting Balance of Diversity Among Major Marine Animal Groups. **Science**, v. 329, n. 5996, p. 1191–1194, 3 set. 2010.

ANDERMANN, T. et al. A Guide to Carrying Out a Phylogenomic Target Sequence Capture Project. **Frontiers in Genetics**, v. 10, p. 1407, 21 fev. 2020.

ARKHIPOVA, I. R. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. **Mobile DNA**, v. 8, n. 1, dez. 2017.

ÁVILA ROBLEDILLO, L. et al. Satellite DNA in *Vicia faba* is characterized by remarkable diversity in its sequence composition, association with centromeres, and replication timing. **Scientific Reports**, v. 8, n. 1, dez. 2018.

BAEZ, M. et al. Analysis of the small chromosomal *Prionium serratum* (Cyperid) demonstrates the importance of reliable methods to differentiate between mono- and holocentricity. **Chromosoma**, v. 129, n. 3–4, p. 285–297, dez. 2020.

BARRETT, C. F. et al. Plastid genomes and deep relationships among the commelinid monocot angiosperms. **Cladistics**, v. 29, n. 1, p. 65–87, fev. 2013.

BEDINI, G.; GARBARI, F.; PERUZZI, L. Karyological knowledge of the Italian vascular flora as inferred by the analysis of "Chrobase.it". **Plant Biosystems - An International Journal Dealing with all Aspects of Plant Biology**, v. 146, n. 4, p. 889–899, dez. 2012.

BELYAYEV, A. et al. Natural History of a Satellite DNA Family: From the Ancestral Genome Component to Species-Specific Sequences, Concerted and Non-Concerted Evolution. **International Journal of Molecular Sciences**, v. 20, n. 5, p. 1201, 9 mar. 2019.

BENNETZEN, J. L.; WANG, H. The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes. **Annual Review of Plant Biology**, v. 65, n. 1, p. 505–530, 29 abr. 2014.

BILINSKI, P. et al. Parallel Altitudinal Clines Reveal Adaptive Evolution Of Genome Size In *Zea mays*. 13 jul. 2017.

BRASLAVSKY, I. et al. Sequence information can be obtained from single DNA molecules. **Proceedings of the National Academy of Sciences**, v. 100, n. 7, p. 3960–3964, 1 abr. 2003.

BRITTEN, R. J.; KOHNE, D. E. Repeated sequences in DNA. **Science**, p. 529–540, 1968.

BROMHAM, L. et al. Exploring the Relationships between Mutation Rates, Life History, Genome Size, Environment, and Species Richness in Flowering Plants. **The American Naturalist**, v. 185, n. 4, p. 507–524, abr. 2015.

BUDDENHAGEN, C. et al. **Anchored Phylogenomics of Angiosperms I: Assessing the Robustness of Phylogenetic Estimates**. [s.l.] Evolutionary Biology, 8 nov. 2016. Disponível em: <http://biorxiv.org/lookup/doi/10.1101/086298>. Acesso em: 13 jul. 2020.

BUDDENHAGEN, C. E. **A view of Rhynchosporeae (Cyperaceae) diversification before and after the application of anchored phylogenomics across the angiosperms**. PhD thesis—Florida, USA: Florida State University, 2016.

BURCHARDT, P. et al. Holocentric Karyotype Evolution in *Rhynchospora* Is Marked by Intense Numerical, Structural, and Genome Size Changes. **Frontiers in Plant Science**, v. 11, p. 536507, 10 set. 2020.

BURGESS, M. B. et al. Effects of apomixis and polyploidy on diversification and geographic distribution in *Amelanchier* (Rosaceae). **American Journal of Botany**, v. 101, n. 8, p. 1375–1387, 2014.

CABRAL, G. et al. Chiasmatic and achiasmatic inverted meiosis of plants with holocentric chromosomes. **Nature Communications**, v. 5, n. 1, dez. 2014.

CACHO, N. I. et al. Genome size evolution is associated with climate seasonality and glucosinolates, but not life history, soil nutrients or range size, across a clade of mustards. **Annals of Botany**, v. 127, n. 7, p. 887–902, 24 jun. 2021.

CAO, Z.; DENG, Z.; MCLAUGHLIN, M. Interspecific genome size and chromosome number variation shed new light on species classification and evolution in caladium. **Journal of the American Society for Horticultural Science**, v. 139, n. 4, p. 449–459, 2014.

ĆERTNER, M. et al. Evolutionary dynamics of mixed-ploidy populations in an annual herb: Dispersal, local persistence and recurrent origins of polyploids. **Annals of Botany**, v. 120, n. 2, p. 303–315, 2017.

CHAFIN, T. K.; DOUGLAS, M. R.; DOUGLAS, M. E. MrBait: universal identification and design of targeted-enrichment capture probes. **Bioinformatics**, v. 34, n. 24, p. 4293–4296, 15 dez. 2018.

CHAN, K. M. A.; MOORE, B. R. SYMMETREE: whole-tree analysis of differential diversification rates. **Bioinformatics**, v. 21, n. 8, p. 1709–1710, 15 abr. 2005.

CHASE, M. et al. Multigene Analyses of Monocot Relationships. **Aliso**, v. 22, n. 1, p. 63–75, 2006.

CHENG, Z. et al. Functional Rice Centromeres Are Marked by a Satellite Repeat and a Centromere-Specific Retrotransposon. **The Plant Cell**, v. 14, n. 8, p. 1691–1704, ago. 2002.

ČÍŽKOVÁ, J. et al. Molecular Analysis and Genomic Organization of Major DNA Satellites in Banana (*Musa* spp.). **PLoS ONE**, v. 8, n. 1, p. e54808, 23 jan. 2013.

CLARKE, L.; CARBON, J. The Structure and Function of Yeast Centromeres. **Annual Review of Genetics**, v. 19, n. 1, p. 29–55, dez. 1985.

COBB, M. Oswald Avery, DNA, and the transformation of biology. **Current Biology**, v. 24, n. 2, p. R55–R60, jan. 2014.

COSTA, L. et al. Comparative cytomolecular analyses reveal karyotype variability related to biogeographic and species richness patterns in Bombacoideae (Malvaceae). **Plant Systematics and Evolution**, v. 303, n. 9, p. 1131–1144, nov. 2017.

COSTA, L. et al. Divide to Conquer: Evolutionary History of Allioideae Tribes (Amaryllidaceae) Is Linked to Distinct Trends of Karyotype Evolution. **Frontiers in Plant Science**, v. 11, p. 320, 7 abr. 2020.

CUACOS, M.; H. FRANKLIN, F. C.; HECKMANN, S. Atypical centromeres in plants—what they can tell us. **Frontiers in Plant Science**, v. 6, 26 out. 2015.

DAVEY, J. W. et al. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. **Nature Reviews Genetics**, v. 12, n. 7, p. 499–510, jul. 2011.

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

DINIZ-FILHO, J. A. F.; DE SANT'ANA, C. E. R.; BINI, L. M. AN EIGENVECTOR METHOD FOR ESTIMATING PHYLOGENETIC INERTIA. **Evolution**, v. 52, n. 5, p. 1247–1262, out. 1998.

DODSWORTH, S. et al. Genomic Repeat Abundances Contain Phylogenetic Signal. **Systematic Biology**, v. 64, n. 1, p. 112–126, 1 jan. 2015.

DODSWORTH, S. Genome skimming for next-generation biodiversity analysis. **Trends in Plant Science**, v. 20, n. 9, p. 525–527, set. 2015.

DODSWORTH, S. et al. Using genomic repeats for phylogenomics: a case study in wild tomatoes (*Solanum* section *Lycopersicon* : Solanaceae). **Biological Journal of the Linnean Society**, v. 117, n. 1, p. 96–105, jan. 2016.

DODSWORTH, S. et al. Genome-wide repeat dynamics reflect phylogenetic distance in closely related allotetraploid *Nicotiana* (Solanaceae). **Plant Systematics and Evolution**, v. 303, n. 8, p. 1013–1020, out. 2017.

DODSWORTH, S. et al. Hyb-Seq for Flowering Plant Systematics. **Trends in Plant Science**, v. 24, n. 10, p. 887–891, out. 2019.

DODSWORTH, S.; CHASE, M. W.; LEITCH, A. R. Is post-polyploidization diploidization the key to the evolutionary success of angiosperms?: Diploidization in Polyploid Angiosperms. **Botanical Journal of the Linnean Society**, v. 180, n. 1, p. 1–5, jan. 2016.

DODSWORTH, S.; LEITCH, A. R.; LEITCH, I. J. Genome size diversity in angiosperms and its influence on gene space. **Current Opinion in Genetics & Development**, v. 35, p. 73–78, dez. 2015.

DRUMMOND, A. J.; RAMBAUT, A. BEAST: Bayesian evolutionary analysis by sampling trees. **BMC Evolutionary Biology**, v. 7, n. 1, p. 214, 2007.

EARNSHAW, W. C. et al. Esperanto for histones: CENP-A, not CenH3, is the centromeric histone H3 variant. **Chromosome Research**, v. 21, n. 2, p. 101–106, abr. 2013.

EDDY, S. R. The C-value paradox, junk DNA and ENCODE. **Current Biology**, v. 22, n. 21, p. R898–R899, nov. 2012.

ELLIOTT, T. A.; GREGORY, T. R. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 370, n. 1678, p. 20140331, 26 set. 2015.

ENKE, N.; FUCHS, J.; GEMEINHOLZER, B. Shrinking genomes? Evidence from genome size variation in *Crepis* (Compositae). **Plant Biology**, v. 13, n. 1, p. 185–193, 2011.

ESCUDEIRO, A. et al. Conservation, Divergence, and Functions of Centromeric Satellite DNA Families in the Bovidae. **Genome Biology and Evolution**, v. 11, n. 4, p. 1152–1165, 1 abr. 2019.

ESCUDERO, M. et al. Genome size stability despite high chromosome number variation in *Carex* gr. *laevigata*. **American Journal of Botany**, v. 102, n. 2, p. 233–238, 2015.

ESCUDERO, M.; WEBER, J. A.; HIPP, A. L. Species coherence in the face of karyotype diversification in holocentric organisms: the case of a cytogenetically variable sedge (*Carex scoparia*, Cyperaceae). **Annals of Botany**, v. 112, n. 3, p. 515–526, ago. 2013.

FELSENSTEIN, J. Phylogenies and the Comparative Method. **The American Naturalist**, v. 125, n. 1, p. 1–15, jan. 1985.

FINNEGAN, D. J. Eukaryotic transposable elements and genome evolution. **Trends in Genetics**, v. 5, p. 103–107, 1989.

FRY, K.; SALSER, W. Nucleotide sequences of HS-α satellite DNA from kangaroo rat dipodomys ordii and characterization of similar sequences in other rodents. **Cell**, v. 12, n. 4, p. 1069–1084, dez. 1977.

GALE, S. Rhynchospora, section Eurhynchospora, in Canada, the United States and the West Indies. In: **Contributions from the Gray Herbarium of Harvard University**. [s.l: s.n.]. v. 151p. 89–iii.

GARCIA, S. et al. Recent updates and developments to plant genome size databases. **Nucleic Acids Research**, v. 42, n. D1, p. D1159–D1166, jan. 2014.

GARRIDO-RAMOS, M. Satellite DNA: An Evolving Topic. **Genes**, v. 8, n. 9, p. 230, 18 set. 2017.

GARRIDO-RAMOS, M. A. Satellite DNA in Plants: More than Just Rubbish. **Cytogenetic and Genome Research**, v. 146, n. 2, p. 153–170, 2015.

GAUT, B. S.; ROSS-IBARRA, J. Selection on Major Components of Angiosperm Genomes. **Science**, v. 320, n. 5875, p. 484–486, 25 abr. 2008.

GEMMELL, N. J. Repetitive DNA: genomic dark matter matters. **Nature Reviews Genetics**, v. 22, n. 6, p. 342–342, jun. 2021.

GLICK, L.; MAYROSE, I. ChromEvol: Assessing the Pattern of Chromosome Number Evolution and the Inference of Polyploidy along a Phylogeny. **Molecular Biology and Evolution**, v. 31, n. 7, p. 1914–1922, jul. 2014.

GNIRKE, A. et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. **Nature Biotechnology**, v. 27, n. 2, p. 182–189, fev. 2009.

GONZÁLEZ, J. et al. Genome-Wide Patterns of Adaptation to Temperate Environments Associated with Transposable Elements in Drosophila. **PLoS Genetics**, v. 6, n. 4, p. e1000905, 8 abr. 2010.

GOOLSBY, E. W. Rapid maximum likelihood ancestral state reconstruction of continuous characters: A rerooting-free algorithm. **Ecology and Evolution**, v. 7, n. 8, p. 2791–2797, abr. 2017.

GOVAERTS, R.; SIMPSON, D. A. (EDS.). **World checklist of Cyperaceae sedges**. Richmond, Surrey: Royal Botanic Gardens, Kew Publ, 2007.

GRAFEN, A. The Phylogenetic Regression. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 326, n. 1233, p. 119–157, 21 dez. 1989.

GREILHUBER, J. et al. Smallest Angiosperm Genomes Found in Lentibulariaceae, with Chromosomes of Bacterial Size. **Plant Biology**, v. 8, n. 6, p. 770–777, nov. 2006.

GROTKOPP, E. et al. EVOLUTION OF GENOME SIZE IN PINES (PINUS) AND ITS LIFE-HISTORY CORRELATES: SUPERTREE ANALYSES. **Evolution**, v. 58, n. 8, p. 1705–1729, ago. 2004.

GUÉNARD, G.; LEGENDRE, P.; PERES-NETO, P. Phylogenetic eigenvector maps: a framework to model and predict species traits. **Methods in Ecology and Evolution**, v. 4, n. 12, p. 1120–1131, dez. 2013.

GUERRA, M. et al. Neocentrics and Holokinetics (Holocentrics): Chromosomes out of the Centromeric Rules. **Cytogenetic and Genome Research**, v. 129, n. 1–3, p. 82–96, 2010.

GUERRA, M. Cytotaxonomy: the end of childhood. **Plant Biosystems**, v. 146, n. 3, p. 703–710, 2012.

GUERRA, M. Agmatoploidy and symploidy: a critical review. **Genetics and Molecular Biology**, v. 39, n. 4, p. 492–496, 27 out. 2016.

GUERRA, M.; RIBEIRO, T.; FELIX, L. P. Monocentric chromosomes in *Juncus* (Juncaceae) and implications for the chromosome evolution of the family. **Botanical Journal of the Linnean Society**, v. 191, n. 4, p. 475–483, 19 nov. 2019.

GUIGNARD, M. S. et al. Genome size and ploidy influence angiosperm species' biomass under nitrogen and phosphorus limitation. **New Phytologist**, v. 210, n. 4, p. 1195–1206, jun. 2016.

HAQUE, F. et al. Solid-state and biological nanopore for real-time sensing of single chemical and sequencing of DNA. **Nano Today**, v. 8, n. 1, p. 56–74, fev. 2013.

HARMON, L. **Phylogenetic Comparative Methods**. [s.l: s.n.].

HARVEY, P. H.; PAGEL, M. D. **The comparative method in evolutionary biology**. Oxford ; New York: Oxford University Press, 1991.

HAVECKER, E. R.; GAO, X.; VOYTAS, D. F. The diversity of LTR retrotransposons. **Genome Biology**, v. 5, n. 6, p. 225, 2004.

HEATHER, J. M.; CHAIN, B. The sequence of sequencers: The history of sequencing DNA. **Genomics**, v. 107, n. 1, p. 1–8, jan. 2016.

HECKMANN, S. et al. Holocentric Chromosomes of *Luzula elegans* Are Characterized by a Longitudinal Centromere Groove, Chromosome Bending, and a Terminal Nucleolus Organizer Region. **Cytogenetic and Genome Research**, v. 134, n. 3, p. 220–228, 2011.

HECKMANN, S. et al. Alternative meiotic chromatid segregation in the holocentric plant *Luzula elegans*. **Nature Communications**, v. 5, n. 1, dez. 2014.

HECKMANN, S.; HOUBEN, A. Holokinetic Centromeres. In: JIANG, J.; BIRCHLER, J. A. (Eds.). . **Plant Centromere Biology**. Oxford, UK: Wiley-Blackwell, 2013. p. 83–94.

HEYDUK, K. et al. Phylogenomic analyses of species relationships in the genus *Sabal* (Arecaceae) using targeted sequence capture. **Biological Journal of the Linnean Society**, v. 117, n. 1, p. 106–120, jan. 2016.

HOF, A. E. VAN'T et al. The industrial melanism mutation in British peppered moths is a transposable element. **Nature**, v. 534, n. 7605, p. 102–105, jun. 2016.

HOUBEN, A.; SCHUBERT, I. DNA and proteins of plant centromeres. **Current Opinion in Plant Biology**, v. 6, n. 6, p. 554–560, dez. 2003.

HUTCHISON, C. A. DNA sequencing: bench to bedside and beyond. **Nucleic Acids Research**, v. 35, n. 18, p. 6227–6237, 28 ago. 2007.

INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM et al. Initial sequencing and analysis of the human genome. **Nature**, v. 409, n. 6822, p. 860–921, 15 fev. 2001.

JAIN, M. et al. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. **Genome Biology**, v. 17, n. 1, p. 239, dez. 2016.

JIAO, Y. et al. Ancestral polyploidy in seed plants and angiosperms. **Nature**, v. 473, n. 7345, p. 97–100, 5 maio 2011.

JOHNSON, M. G. et al. A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering. **Systematic Biology**, v. 68, n. 4, p. 594–606, 1 jul. 2019.

KAISER, A. D.; WU, R. Structure and Function of DNA Cohesive Ends. **Cold Spring Harbor Symposia on Quantitative Biology**, v. 33, n. 0, p. 729–734, 1 jan. 1968.

KALENDAR, R. et al. Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. **Proceedings of the National Academy of Sciences**, v. 97, n. 12, p. 6603–6607, 6 jun. 2000.

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

KAPITONOV, V. V.; JURKA, J. A universal classification of eukaryotic transposable elements implemented in Repbase. **Nature Reviews Genetics**, v. 9, n. 5, p. 411–412, maio 2008.

KELLY, L. J. et al. Analysis of the giant genomes of *Fritillaria* (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. **New Phytologist**, v. 208, n. 2, p. 596–607, out. 2015.

KIMURA, Y. et al. OARE-1, a Ty1-copia Retrotransposon in Oat Activated by Abiotic and Biotic Stresses. **Plant and Cell Physiology**, v. 42, n. 12, p. 1345–1354, 15 dez. 2001.

KOO, D.-H. et al. Rapid divergence of repetitive DNAs in *Brassica* relatives. **Genomics**, v. 97, n. 3, p. 173–185, mar. 2011.

KRAAIJEVELD, K. Genome Size and Species Diversification. **Evolutionary Biology**, v. 37, n. 4, p. 227–233, dez. 2010.

KÜKENTHAL, G. Vorarbeiten zu einer Monographie der Rhynchosporoideae. **Repertorium novarum specierum regni vegetabilis**, v. 47, n. 15–20, p. 209–216, 1939.

LANDIS, J. B. et al. Impact of whole-genome duplication events on diversification rates in angiosperms. **American Journal of Botany**, v. 105, n. 3, p. 348–363, mar. 2018.

LANG, D. et al. Genome-Wide Phylogenetic Comparative Analysis of Plant Transcriptional Regulation: A Timeline of Loss, Gain, Expansion, and Correlation with Complexity. **Genome Biology and Evolution**, v. 2, p. 488–503, 1 jan. 2010.

LEE, H.-R. et al. From The Cover: Chromatin immunoprecipitation cloning reveals rapid evolutionary patterns of centromeric DNA in *Oryza* species. **Proceedings of the National Academy of Sciences**, v. 102, n. 33, p. 11793–11798, 16 ago. 2005.

LEMOPOULOS, A. et al. Comparing RADseq and microsatellites for estimating genetic diversity and relatedness — Implications for brown trout conservation. **Ecology and Evolution**, v. 9, n. 4, p. 2106–2120, fev. 2019.

LEVENE, M. J. et al. Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. **Science**, v. 299, n. 5607, p. 682–686, 31 jan. 2003.

LEVIN, S. A. **The princeton guide to ecology**. Princeton: Princeton University Press, 2012.

LIM, K. G. et al. Review of tandem repeat search tools: a systematic approach to evaluating algorithmic performance. **Briefings in Bioinformatics**, v. 14, n. 1, p. 67–81, 1 jan. 2013.

LINDER, H. P.; RUDALL, P. J. Evolutionary History of Poales. **Annual Review of Ecology, Evolution, and Systematics**, v. 36, n. 1, p. 107–124, 1 dez. 2005.

LOWER, S. S. et al. Genome Size in North American Fireflies: Substantial Variation Likely Driven by Neutral Processes. **Genome biology and evolution**, v. 9, n. 6, p. 1499–1512, 2017.

LUKHTANOV, V. A. et al. Versatility of multivalent orientation, inverted meiosis, and rescued fitness in holocentric chromosomal hybrids. **Proceedings of the National Academy of Sciences**, v. 115, n. 41, p. E9610–E9619, 9 out. 2018.

LYU, H. 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, jan. 2018.

MACAS, J. et al. In Depth Characterization of Repetitive DNA in 23 Plant Genomes Reveals Sources of Genome Size Variation in the Legume Tribe Fabeae. **PLOS ONE**, v. 10, n. 11, p. e0143424, 25 nov. 2015.

MACAS, J.; MESZAROS, T.; NOUZOVA, M. PlantSat: a specialized database for plant satellite repeats. **Bioinformatics**, v. 18, n. 1, p. 28–35, 1 jan. 2002.

MANDEL, J. R. et al. A Target Enrichment Method for Gathering Phylogenetic Information from Hundreds of Loci: An Example from the Compositae. **Applications in Plant Sciences**, v. 2, n. 2, p. 1300085, fev. 2014.

MARGUERAT, S.; BÄHLER, J. RNA-seq: from technology to biology. **Cellular and Molecular Life Sciences**, v. 67, n. 4, p. 569–579, fev. 2010.

MARGULIES, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. **Nature**, v. 437, n. 7057, p. 376–380, 15 set. 2005.

MARKOS, S.; BALDWIN, B. G. Higher-Level Relationships and Major Lineages of Lessingia (Compositae, Astereae) Based on Nuclear rDNA Internal and External Transcribed Spacer (ITS and ETS) Sequences. **Systematic Botany**, v. 26, n. 1, p. 168–183, 2001.

MARQUES, A. et al. Holocentromeres in *Rhynchospora* are associated with genome-wide centromere-specific repeat arrays interspersed among euchromatin. **Proceedings of the National Academy of Sciences of the United States of America**, v. 112, n. 44, p. 13633–13638, 2015.

MARQUES, A. M. et al. Refinement of the karyological aspects of *Psidium guineense* (Swartz, 1788): A comparison with Psidium guajava (Linnaeus, 1753). **Comparative Cytogenetics**, v. 10, n. 1, p. 117–128, 2016.

MÁRQUEZ-CORRO, J. I.; ESCUDERO, M.; LUCEÑO, M. Do holocentric chromosomes represent an evolutionary advantage? A study of paired analyses of diversification rates of lineages with holocentric chromosomes and their monocentric closest relatives. **Chromosome Research**, v. 26, n. 3, p. 139–152, set. 2018.

MATSUNAGA, W. et al. A small RNA mediated regulation of a stress-activated retrotransposon and the tissue specific transposition during the reproductive period in Arabidopsis. **Frontiers in Plant Science**, v. 6, 9 fev. 2015.

MATSUNO, M. et al. Evolution of a Novel Phenolic Pathway for Pollen Development. **Science**, v. 325, n. 5948, p. 1688–1692, 25 set. 2009.

MCCLINTOCK, B. Mutable loci in maize. **Carnegie Institution of Washington Yearbook**, v. 47, p. 155–169, 1948.

MELTERS, D. P. et al. Holocentric chromosomes: convergent evolution, meiotic adaptations, and genomic analysis. **Chromosome Research**, v. 20, n. 5, p. 579–593, jul. 2012.

MOORE, L. L.; MORRISON, M.; ROTH, M. B. HCP-1, a protein involved in chromosome segregation, is localized to the centromere of mitotic chromosomes in *Caenorhabditis elegans*. **The Journal of Cell Biology**, v. 147, n. 3, p. 471–480, 1 nov. 1999.

MRAVINAC, B.; PLOHL, M.; UGARKOVIĆ, Đ. Preservation and High Sequence Conservation of Satellite DNAs Suggest Functional Constraints. **Journal of Molecular Evolution**, v. 61, n. 4, p. 542–550, out. 2005.

NAGAKI, K. et al. Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key functional DNA element of Arabidopsis thaliana centromeres. **Genetics**, v. 163, n. 3, p. 1221–1225, mar. 2003.

NAGAKI, K.; MURATA, M. Characterization of CENH3 and centromere-associated DNA sequences in sugarcane. **Chromosome Research**, v. 13, n. 2, p. 195–203, fev. 2005.

NEUMANN, P. et al. Plant centromeric retrotransposons: a structural and cytogenetic perspective. **Mobile DNA**, v. 2, n. 1, p. 4, 2011.

NEUMANN, P. 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, dez. 2019.

NOVAK, P. 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, 15 mar. 2013.

OAKLEY, T. H. et al. Comparative Methods for the Analysis of Gene-Expression Evolution: An Example Using Yeast Functional Genomic Data. **Molecular Biology and Evolution**, v. 22, n. 1, p. 40–50, jan. 2005.

OGUTCEN, E. et al. Phylogenomics of Gesneriaceae using targeted capture of nuclear genes. **Molecular Phylogenetics and Evolution**, v. 157, p. 107068, abr. 2021.

O'MEARA, B. C. Evolutionary Inferences from Phylogenies: A Review of Methods. **Annual Review of Ecology, Evolution, and Systematics**, v. 43, n. 1, p. 267–285, dez. 2012.

O'MEARA, B. C. et al. Non-equilibrium dynamics and floral trait interactions shape extant angiosperm diversity. **Proceedings of the Royal Society B: Biological Sciences**, v. 283, n. 1830, p. 20152304, 11 maio 2016.

PANDIT, M. K.; WHITE, S. M.; POCOCK, M. J. O. The contrasting effects of genome size, chromosome number and ploidy level on plant invasiveness: a global analysis. **New Phytologist**, v. 203, n. 2, p. 697–703, jul. 2014.

PELLICER, J.; FAY, M. F.; LEITCH, I. J. The largest eukaryotic genome of them all?: THE LARGEST EUKARYOTIC GENOME? **Botanical Journal of the Linnean Society**, v. 164, n. 1, p. 10–15, 15 set. 2010.

PENNELL, M. W.; HARMON, L. J. An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology: Integrative comparative methods. **Annals of the New York Academy of Sciences**, v. 1289, n. 1, p. 90–105, jun. 2013.

PERUZZI, L. et al. Does actually mean chromosome number increase with latitude in vascular plants? An answer from the comparison of Italian, Slovak and Polish floras. **Comparative Cytogenetics**, v. 6, n. 4, p. 371–377, 19 nov. 2012.

PESKA, V.; GARCIA, S. Origin, Diversity, and Evolution of Telomere Sequences in Plants. **Frontiers in Plant Science**, v. 11, p. 117, 21 fev. 2020.

PETRACCIOLI, A. et al. A novel satellite DNA isolated in *Pecten jacobaeus* shows high sequence similarity among molluscs. **Molecular Genetics and Genomics**, v. 290, n. 5, p. 1717–1725, out. 2015.

PIEGU, B. et al. Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. **Genome Research**, v. 16, n. 10, p. 1262–1269, 1 out. 2006.

PIÉGU, B. et al. A survey of transposable element classification systems – A call for a fundamental update to meet the challenge of their diversity and complexity. **Molecular Phylogenetics and Evolution**, v. 86, p. 90–109, maio 2015.

PLOHL, M. et al. Satellite DNAs between selfishness and functionality: Structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. **Gene**, v. 409, n. 1–2, p. 72–82, fev. 2008.

PLOHL, M.; MEŠTROVIĆ, N.; MRAVINAC, B. Centromere identity from the DNA point of view. **Chromosoma**, v. 123, n. 4, p. 313–325, ago. 2014.

PLOHL, M.; MEš TROVIC, N.; MRAVINAC, B. Satellite DNA Evolution. In: GARRIDO-RAMOS, M. A. (Ed.). . **Genome Dynamics**. Basel: S. KARGER AG, 2012. v. 7p. 126–152.

PUTTICK, M. N.; CLARK, J.; DONOGHUE, P. C. J. Size is not everything: rates of genome size evolution, not *C* -value, correlate with speciation in angiosperms. **Proceedings of the Royal Society B: Biological Sciences**, v. 282, n. 1820, p. 20152289, 7 dez. 2015.

RABOSKY, D. L. Automatic Detection of Key Innovations, Rate Shifts, and Diversity-Dependence on Phylogenetic Trees. **PLoS ONE**, v. 9, n. 2, p. e89543, 26 fev. 2014.

RAMALLO, E. et al. Reme1, a Copia retrotransposon in melon, is transcriptionally induced by UV light. **Plant Molecular Biology**, v. 66, n. 1–2, p. 137–150, jan. 2008.

RAYBURN, A. L. et al. C-Band Heterochromatin and DNA Content in *Zea mays*. **American Journal of Botany**, v. 72, n. 10, p. 1610, out. 1985.

RIBEIRO, T. et al. Centromeric and non-centromeric satellite DNA organisation differs in holocentric *Rhynchospora* species. **Chromosoma**, v. 126, n. 2, p. 325–335, mar. 2017.

RIBEIRO, T. et al. Are holocentrics doomed to change? Limited chromosome number variation in *Rhynchospora* Vahl (Cyperaceae). **Protoplasma**, v. 255, n. 1, p. 263–272, 2018.

RICHARD, G.-F.; KERREST, A.; DUJON, B. Comparative Genomics and Molecular Dynamics of DNA Repeats in Eukaryotes. **Microbiology and Molecular Biology Reviews**, v. 72, n. 4, p. 686–727, 1 dez. 2008.

ROA, F.; GUERRA, M. Distribution of 45S rDNA sites in chromosomes of plants: Structural and evolutionary implications. **BMC Evolutionary Biology**, v. 12, n. 1, p. 225, 2012.

ROALSON, E. H. A Synopsis of Chromosome Number Variation in the Cyperaceae. **The Botanical Review**, v. 74, n. 2, p. 209–393, jun. 2008.

ROALSON, E.; MCCUBBIN, A.; WHITKUS, R. Chromosome Evolution in Cyperales. **Aliso**, v. 23, n. 1, p. 62–71, 2007.

ROBLES, F. et al. Evolution of ancient satellite DNAs in sturgeon genomes. **Gene**, v. 338, n. 1, p. 133–142, ago. 2004.

RONAGHI, M. et al. Real-Time DNA Sequencing Using Detection of Pyrophosphate Release. **Analytical Biochemistry**, v. 242, n. 1, p. 84–89, nov. 1996.

RONQUIST, F.; HUELSENBECK, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. **Bioinformatics**, v. 19, n. 12, p. 1572–1574, 12 ago. 2003.

RUIZ-RUANO, F. J. et al. High-throughput analysis of the satellitome illuminates satellite DNA evolution. **Scientific Reports**, v. 6, n. 1, set. 2016.

SABATH, N. et al. Dioecy does not consistently accelerate or slow lineage diversification across multiple genera of angiosperms. **New Phytologist**, v. 209, n. 3, p. 1290–1300, fev. 2016.

SADER, M. A. et al. The role of chromosome changes in the diversification of *Passiflora* L. (Passifloraceae). **Systematics and Biodiversity**, v. 17, n. 1, p. 7–21, 2 jan. 2019.

SALSER, W. et al. Investigation of the organization of mammalian chromosomes at the DNA sequence level. **Federation Proceedings**, v. 35, n. 1, p. 23–35, jan. 1976.

SANGER, F. The terminal peptides of insulin. **Biochemical Journal**, v. 45, n. 5, p. 563–574, 1 jan. 1949.

SANGER, F.; COULSON, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. **Journal of Molecular Biology**, v. 94, n. 3, p. 441–448, maio 1975.

SANGER, F.; TUPPY, H. The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. **Biochemical Journal**, v. 49, n. 4, p. 481–490, 1 set. 1951.

SARGENT, R. D. Floral symmetry affects speciation rates in angiosperms. **Proceedings of the Royal Society of London. Series B: Biological Sciences**, v. 271, n. 1539, p. 603–608, 22 mar. 2004.

SCHLEY, R. J. et al. **The Ecology of Palm Genomes: Repeat-associated genome size expansion is constrained by aridity**. [s.l.] Evolutionary Biology, 8 nov. 2021. Disponível em: <http://biorxiv.org/lookup/doi/10.1101/2021.11.04.467295>. Acesso em: 13 nov. 2021.

SCHMICKL, R. et al. Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African *Oxalis* (Oxalidaceae). **Molecular Ecology Resources**, v. 16, n. 5, p. 1124–1135, set. 2016.

SCHMUTHS, H. Genome Size Variation among Accessions of Arabidopsis thaliana. **Annals of Botany**, v. 93, n. 3, p. 317–321, 26 jan. 2004.

SCHRADER, L. et al. Transposable element islands facilitate adaptation to novel environments in an invasive species. **Nature Communications**, v. 5, n. 1, dez. 2014.

SCHRADER, L.; SCHMITZ, J. The impact of transposable elements in adaptive evolution. **Molecular Ecology**, v. 28, n. 6, p. 1537–1549, mar. 2019.

SINISCALCHI, C. M. et al. Phylogenomics Yields New Insight Into Relationships Within Vernonieae (Asteraceae). **Frontiers in Plant Science**, v. 10, p. 1224, 17 out. 2019.

SLOWINSKI, J. B.; GUYER, C. Testing Whether Certain Traits have Caused Amplified Diversification: An Improved Method Based on a Model of Random Speciation and Extinction. **The American Naturalist**, v. 142, n. 6, p. 1019–1024, dez. 1993.

SMITH, L. M. et al. Fluorescence detection in automated DNA sequence analysis. **Nature**, v. 321, n. 6071, p. 674–679, jun. 1986.

SOUZA, G. 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, jun. 2019.

SPALINK, D. et al. Biogeography of the cosmopolitan sedges (Cyperaceae) and the area-richness correlation in plants. **Journal of Biogeography**, v. 43, n. 10, p. 1893–1904, out. 2016.

SPROUL, J. S.; BARTON, L. M.; MADDISON, D. R. Repetitive DNA Profiles Reveal Evidence of Rapid Genome Evolution and Reflect Species Boundaries in Ground Beetles. **Systematic Biology**, v. 69, n. 6, p. 1137–1148, 1 nov. 2020.

STAMATAKIS, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. **Bioinformatics**, v. 22, n. 21, p. 2688–2690, 1 nov. 2006.

STEINER, F. A.; HENIKOFF, S. Holocentromeres are dispersed point centromeres localized at transcription factor hotspots. **eLife**, v. 3, 8 abr. 2014.

STRAUB, S. C. K. 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, fev. 2012.

SUN, X.; WAHLSTROM, J.; KARPEN, G. Molecular structure of a functional *Drosophila* centromere. **Cell**, v. 91, n. 7, p. 1007–1019, 26 dez. 1997.

SUN, Y. et al. Twenty years of plant genome sequencing: achievements and challenges. **Trends in Plant Science**, p. S1360138521002818, nov. 2021.

SWIFT, H. H. The desoxyribose nucleic acid content of animal nuclei. **Physiological Zoology**, v. 23, n. 3, p. 169–198, jul. 1950.

TANK, D. C. et al. Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. **New Phytologist**, v. 207, n. 2, p. 454–467, jul. 2015.

THE *C. ELEGANS* SEQUENCING CONSORTIUM. Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. **Science**, v. 282, n. 5396, p. 2012–2018, 11 dez. 1998.

THOMAS, C. A. The Genetic Organization of Chromosomes. **Annual Review of Genetics**, v. 5, n. 1, p. 237–256, dez. 1971.

THOMAS, W. W. Two new species of *Rhynchospora* (Cyperaceae) from Bahia, Brazil, and new combinations in *Rhynchospora* section *Pleurostachys*. **Brittonia**, v. 72, n. 3, p. 273–281, set. 2020.

THOMAS, WM. W.; ARAÚJO, A. C.; ALVES, M. V. A Preliminary Molecular Phylogeny of the Rhynchosporeae (Cyperaceae). **The Botanical Review**, v. 75, n. 1, p. 22–29, fev. 2009.

TSOUMANI, K. T. et al. Molecular Characterization and Chromosomal Distribution of a Species-Specific Transcribed Centromeric Satellite Repeat from the Olive Fruit Fly, *Bactrocera oleae*. **PLoS ONE**, v. 8, n. 11, p. e79393, 14 nov. 2013.

UYEDA, J. C.; ZENIL-FERGUSON, R.; PENNELL, M. W. Rethinking phylogenetic comparative methods. **Systematic Biology**, v. 67, n. 6, p. 1091–1109, 1 nov. 2018.

VAN DE PAER, C. et al. Mitogenomics of *Hesperelaea*, an extinct genus of Oleaceae. **Gene**, v. 594, n. 2, p. 197–202, dez. 2016.

VAN DIJK, E. L. et al. Ten years of next-generation sequencing technology. **Trends in Genetics**, v. 30, n. 9, p. 418–426, set. 2014.

VAN-LUME, B. 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, dez. 2017.

VANZELA, A. L. L.; GUERRA, M. Heterochromatin differentiation in holocentric chromosomes of *Rhynchospora* (Cyperaceae). **Genetics and Molecular Biology**, v. 23, n. 2, p. 453–456, jun. 2000.

VENTER, J. C. et al. The Sequence of the Human Genome. **Science**, v. 291, n. 5507, p. 1304–1351, 16 fev. 2001.

VINOGRADOV, A. E. Selfish DNA is maladaptive: evidence from the plant Red List. **Trends in Genetics**, v. 19, n. 11, p. 609–614, nov. 2003.

VITALES, D.; GARCIA, S.; DODSWORTH, S. Reconstructing phylogenetic relationships based on repeat sequence similarities. **Molecular Phylogenetics and Evolution**, v. 147, p. 106766, jun. 2020.

VOELKERDING, K. V.; DAMES, S. A.; DURTSCHI, J. D. Next-Generation Sequencing: From Basic Research to Diagnostics. **Clinical Chemistry**, v. 55, n. 4, p. 641–658, 1 abr. 2009.

WANG, H.-J. et al. Resolving interspecific relationships within evolutionarily young lineages using RNA-seq data: An example from Pedicularis section Cyathophora (Orobanchaceae). **Molecular Phylogenetics and Evolution**, v. 107, p. 345–355, fev. 2017.

WATSON, J. D.; CRICK, F. H. C. A structure for Deoxyribose Nucleic Acid. **Nature**, v. 171, n. 4356, p. 737–738, 1953.

WEISS-SCHNEEWEISS, H. et al. Exploring the repeats' landscape and its impact on genome evolution and plant diversification. In: Germany: Koeltz Scientific Books, 2015.

WEISS-SCHNEEWEISS, H.; SCHNEEWEISS, G. M. Karyotype Diversity and Evolutionary Trends in Angiosperms. In: GREILHUBER, J.; DOLEZEL, J.; WENDEL, J. F. (Eds.). . **Plant Genome Diversity Volume 2**. Vienna: Springer Vienna, 2013. p. 209–230.

WEITEMIER, K. et al. Hyb-Seq: Combining Target Enrichment and Genome Skimming for Plant Phylogenomics. **Applications in Plant Sciences**, v. 2, n. 9, p. 1400042, set. 2014.

WICKER, T. et al. A unified classification system for eukaryotic transposable elements. **Nature Reviews Genetics**, v. 8, n. 12, p. 973–982, dez. 2007.

WILLARD, H. F.; WAYE, J. S. Chromosome-specific subsets of human alpha satellite DNA: analysis of sequence divergence within and between chromosomal subsets and evidence for an ancestral pentameric repeat. **Journal of Molecular Evolution**, v. 25, n. 3, p. 207–214, 1987.

WU, R.; KAISER, A. D. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. **Journal of Molecular Biology**, v. 35, n. 3, p. 523–537, jan. 1968.

YODER, J. B. et al. Phylogenetic Signal Variation in the Genomes of *Medicago* (Fabaceae). **Systematic Biology**, v. 62, n. 3, p. 424–438, 1 maio 2013.

ZAITLIN, D.; PIERCE, A. Nuclear DNA content in *Sinningia* (Gesneriaceae); Intraspecific genome size variation and genome characterization in S. speciosa. **Genome**, v. 53, n. 12, p. 1066–1082, 2010.

ZALLEN, D. T. Despite Franklin’s work, Wilkins earned his Nobel. **Nature**, v. 425, n. 6953, p. 15–15, set. 2003.

ZEDEK, F.; BUREŠ, P. Holocentric chromosomes: from tolerance to fragmentation to colonization of the land. **Annals of Botany**, v. 121, n. 1, p. 9–16, 25 jan. 2018.

ZHONG, C. X. et al. Centromeric Retroelements and Satellites Interact with Maize Kinetochore Protein CENH3. **The Plant Cell**, v. 14, n. 11, p. 2825–2836, nov. 2002.

APÊNDICE A – AIMING OFF THE TARGET: RECYCLING TARGET CAPTURE SEQUENCING READS FOR INVESTIGATING REPETITIVE DNA

*Original paper published at Annals of Botany (https://doi.org/10.1093/aob/mcab063)

Aiming off the target: recycling target capture sequencing reads for investigating repetitive DNA

Lucas Costa[1], André Marques[2], Chris Buddenhagen[3], William Wayt Thomas[4], Bruno Huettel[5] Veit Schubert[6], Steven Dodsworth[7], Andreas Houben[6], Gustavo Souza[1], Andrea Pedrosa-Harand[1*]

[1] *Laboratory of Plant Cytogenetics and Evolution, Department of Botany, Federal University of Pernambuco, Recife-PE, Brazil*

[2] *Max Planck Institute for Plant Breeding Research, Cologne, Germany*

[3] *AgResearch, Plant Functional Biology, Ruakura, New Zealand*

[4] *New York Botanical Garden, Bronx, New York, United States of America*

[5] *Max Planck Genome Centre Cologne, Max Planck Institute for Plant Breeding Research, Cologne, Germany*

[6] *Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Seeland, Germany*

[7] *School of Life Sciences, University of Bedfordshire, Luton, UK*

* andrea.harand@ufpe.br

Running Title: Target capture sequencing for investigating repetitive DNA

SUMMARY

- *Background and Aims*

  With the advance of high-throughput sequencing (HTS), reduced-representation methods such as target capture sequencing (TCS) emerged as cost-efficient ways of gathering genomic information, particularly from coding regions. As the off-target reads from such sequencing are expected to be similar to genome skimming (GS), we assessed the quality of repeat characterization in plant genomes using this data.

- *Methods*

  Repeat composition obtained from TCS datasets of five *Rhynchospora* (Cyperaceae) species were compared with GS data from the same taxa. In addition, a FISH probe was designed based on the most abundant satellite found in the TCS dataset of *Rhynchospora cephalotes*. Finally, repeat-based phylogenies of the five *Rhynchospora* species were constructed based on the GS and TCS dataset and the topologies were compared with a gene-alignment based phylogenetic tree.

- *Key Results*

  All the major repetitive DNA families were identified in TCS, including repeats that showed abundances as low as 0.01% in the GS data. Rank correlation between GS and TCS repeat abundances were moderately high (r = 0.58-0.85), increasing after filtering out the targeted loci from the raw TCS reads (r = 0.66-0.92). Repeat data obtained by TCS was also reliable to develop a cytogenetic probe of a new variant of the holocentromeric satellite *Tyba*. Repeat-based phylogenies from TCS data were congruent with those obtained from GS data and the gene-alignment tree.

- *Conclusions*

  Our results show that off-target TCS reads can be recycled to identify repeats for cyto- and phylogenomic investigations. Given the growing availability of TCS reads, driven

by global phylogenomic projects, our strategy represents a way to recycle genomic data and contribute to a better characterization of plant biodiversity.

Keywords: Genome Skimming, Holocentric, Reduced-representation sequencing, RepeatExplorer, *Rhynchospora*, Satellite DNA, Transposable elements

INTRODUCTION

One intriguing aspect of plant genomes is the staggering 2,400-fold variation in DNA content among species (Pellicer *et al.*, 2010). Advances in genomics have led to the discovery that most of this diversity is the result of variable amounts of repetitive DNA, commonly divided into tandem repeats and dispersed repeats (Weiss-Schneeweiss *et al.*, 2015; Elliott and Gregory, 2015). Tandemly distributed satellite DNAs are remarkable for their fast evolution and variance in abundance and structure at all hierarchical levels (Novák *et al.*, 2017; Ávila Robledillo *et al.*, 2018). As for dispersed repetitive sequences, transposable elements (TE) are especially abundant in flowering plant genomes (Jurka *et al.*, 2011; Galindo-González *et al.* 2017). Retrotransposons, particularly the ones possessing a long terminal repeat (LTR-retrotransposons) account for most of this abundance (Weiss-Schneeweiss *et al.*, 2015) with two major superfamilies being recognized (Ty1/copia and Ty3/gypsy) based on the order of the protein coding domains, and further divided into a number of lineages according to phylogenetic distance (Neumann *et al.*, 2019).

Contrasting with the previous notion that repetitive DNA was no more than "junk DNA", cytogenomic studies in the last few decades helped to uncover possible roles for tandem and dispersed repeats. Specific satellite DNAs and retrotransposons have been found to have a functional and/or structural role in centromeres (Cheng *et al.*, 2002; Nagaki *et al.*, 2003; Houben and Schubert, 2003; Marques *et al.*, 2015; Macas *et al.*, 2015; Ribeiro *et al.*, 2017). The role of LTR-retrotransposons in genome size variation led to investigations correlating these to heterochromatin distribution, ecological variables, community structure and plant distribution (Guignard *et al.*, 2016; Van-Lume *et al.*, 2017; Lyu *et al.*, 2018; Souza *et al.*, 2019). The activity of transposable elements in the host genome can also cause modifications to gene regulation and the formation of retrogenes, generating morphological innovations and impacting speciation processes (Schrader and Schmitz, 2019). Moreover, lineage-specific satellite DNAs have been widely used as efficient cytomolecular markers, allowing the identification of

chromosome pairs and elucidating chromosome rearrangement and duplication events (Koo *et al.*, 2011; Čížková *et al.*, 2013; Ávila Robledillo *et al.*, 2018; Ribeiro *et al.*, 2020) .

The fast evolution of repetitive DNA, with many satellite families being genus- or species-specific, impair their use in phylogenetic studies, as concerted evolution and homogenisation reduces sequence variability for comparative studies across taxa (Macas *et al.*, 2015; Mascagni *et al.*, 2020; Ribeiro *et al.*, 2020). Their abundances, however, have phylogenetic significance, as demonstrated by a method to reconstruct phylogenetic relationships using the abundance of different repetitive elements (Dodsworth *et al.*, 2015). This method has proven useful to elucidate relationships in different groups of plants and animals (Dodsworth *et al.*, 2017; Bolsheva *et al.*, 2019; Martín-Peciña *et al.*, 2019). Other methods have assessed the usefulness of repeat-based phylogenetic analysis. More recently, it was demonstrated that sequence similarity measures of repeated sequences can also be used as characters to resolve phylogenetic relationships (Vitales *et al.*, 2020). In a similar approach, assembly and alignment free (AAF) methods can be applied to high complexity fractions of the genome, such as repetitive DNA, in a phylogenomic framework (Fan *et al.*, 2015; Sarmashghi *et al.*, 2019).

In order to identify and characterize the diversity of repeats in a genome, the RepeatExplorer pipeline was created, using a graph-based clustering algorithm to group high-copy sequences based on similarity, with posterior identification of protein-coding domains by cross-referencing with an extensive group of up-to-date databases (Novak *et al.*, 2013). RepeatExplorer can identify repetitive elements with approximately 0.1× genome coverage, most commonly known as genome skimming (GS), which is often sufficient to study high-copy nuclear and organellar DNA (Straub *et al.*, 2012; Dodsworth, 2015; Dodsworth *et al.*, 2019). The GS method is just one of several “reduced representation” methods of high-throughput sequencing. Throughout the last decade, a number of these sequencing methods have been

proposed, generating high quality sequencing data at decreasing costs, such as restriction site-associated sequencing (RAD-seq, Eaton *et al.*, 2016) and transcriptome sequencing (RNA-seq, Wang *et al.*, 2017).

Another widely used reduced representation method is target capture sequencing (TCS), in which several genomic probes are designed to "capture" and enrich specific low-copy coding regions of the nuclear genome (Albert *et al.*, 2007; Gnirke *et al.*, 2009). These probes can be designed based on conserved regions retrieved from the alignment of several genomes of a divergent set of organisms (e.g. Anchored Hybrid Enrichment, Lemmon *et al.*, 2012) or by comparing transcriptomic data to search for a conserved set of orthologs across a group (Johnson *et al.*, 2019). Since the development of TCS (Albert *et al.*, 2007), many sets of probes have been developed and applied in both plants and animals (Cosart *et al.*, 2011; Faircloth *et al.*, 2012; Mandel *et al.*, 2014; Ilves and López-Fernández, 2014; Sass *et al.*, 2016; Heyduk *et al.*, 2016; Schmickl *et al.*, 2016). Moreover, the nature of these probe design approaches have allowed the development of universal probe sets, with the potential to be used across all of the angiosperms (e.g. Buddenhagen *et al.*, 2016; Johnson *et al.*, 2019). In addition to the advantage of being universal, high recovery rate of target regions (or the number of enriched targeted loci in the final library) is often achievable with very low "enrichment efficiency" (percentage of library reads successfully mapped to a target sequence) (Johnson *et al.*, 2019), meaning that a TCS enriched library will often contain a high number of "off-target" reads. The use of these off-target reads, in combination with the enriched reads, has been referred to as "Hyb-Seq" (hybrid sequencing, Weitemier *et al.*, 2014). As the off-target reads are often rich in high-copy DNA, Hyb-Seq approaches have been used to reconstruct whole plastomes and to obtain ribosomal DNA profiles (Weitemier *et al.*, 2014; Schmickl *et al.*, 2016; Sproul *et al.*, 2020).

To check whether off-target reads could also be recycled for identifying, and potentially quantifying repetitive DNA, we selected the sedge *Rhynchospora* Vahl as a model.

Rhynchospora is one of the largest genera of Cyperaceae Juss., comprising approximately 350 species, but it is under-studied from a phylogenetic point of view (Thomas *et al.*, 2009; Buddenhagen *et al.*, 2016). Cytologically, *Rhynchospora* has been of great interest due to its holocentric chromosomes, which present the centromere dispersed along the sister chromatids rather than localized in a primary constriction (Bureš *et al.*, 2013). Moreover, cytogenomic studies on *R. pubera* (Vahl) Boeckeler have led to the discovery of the first centromere-specific satellite DNA reported in a holocentric organism, Tyba (Marques *et al.*, 2015). Subsequent studies confirmed the presence of Tyba in other *Rhynchospora* species (Ribeiro *et al.*, 2017). Nevertheless, other non-centromeric satellites found in *Rhynchospora* species showed the typical block-like pattern on localized chromosomal regions (Ribeiro *et al.*, 2017).

Large-scale repeat analysis covering all major clades of *Rhynchospora* would provide valuable insights into the repeat evolution of this genus. Recently, using a set of probes developed by Anchored Hybrid Enrichment (Buddenhagen *et al.*, 2016), a number of *Rhynchospora* species were sequenced using a TCS approach. Here, we assessed the quality of repeat characterization in five *Rhynchospora* species (of which we also possessed GS data) from this dataset. As a considerable part of the off-target reads can be sequences close to the original targeted loci (Dodsworth *et al.*, 2019), we searched for sequencing bias by comparing the target results with results obtained from genome skimming (i.e. unenriched libraries). We specifically addressed three questions: 1) Can we characterize the repetitive DNA fraction of *Rhynchospora* genomes using TCS data, compared to GS data?; 2) Can we develop cytological markers from TCS clustering data?; and 3) Can we use TCS data to reconstruct repeat-based phylogenetic trees?

MATERIALS AND METHODS

Plant material and sequence data

To assess whether the off-target reads from TCS could be used in a similar manner as GS data to identify high-copy repeats, a comparison between those data types was necessary. At first, only three species fitted this requirement (*R. globosa*, *R. pubera* and *R. tenuis*), thus two additional species were collected for GS sequencing. Individuals of *Rhynchospora cephalotes* (L.) Vahl and *R. exaltata* Kunth were collected near the towns of Jacaraú (Paraíba, Brazil, voucher UFP87625) and Jaqueira (Pernambuco, Brazil, voucher JPB51537), respectively. These individuals were cultivated in (i) the experimental garden of the Laboratory of Plant Cytogenetic and Evolution at the Universidade Federal de Pernambuco (Brazil), (ii) the greenhouse of the Max Plank Institute for Plant Breeding Research (Cologne, Germany) and (iii) the greenhouse of the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK Gatersleben, Germany).

We downloaded available short read archive data of *R. pubera* (Vahl) Boeckeler (Marques *et al*., 2015, BioProject PRJEB9643) from the NCBI GenBank (www.ncbi.nlm.nih.gov). Genome skimming sequences of *R. globosa* (Kunth) Roem. & Schult. and *R. tenuis* Link were obtained from Ribeiro *et al*. (2017) and deposited on GenBank under BioProject PRJNA672922. TCS reads (150 bp) for *R. cephalotes*, *R. exaltata*, *R. globosa*, *R. pubera* and *R. tenuis* were obtained from Buddenhagen (2016) and deposited on GenBank under BioProject PRJNA672127.

DNA extraction and sequencing

Genomic DNA from *R. cephalotes* and *R. exaltata* was extracted from leaves with NucleoBond HMW DNA kit (Macherey and Nagel, Düren, Germany). Quality was assessed with Agilent TapeStation and the gDNA was quantified by Qubit BR assay (Thermo). An Illumina-compatible library was then prepared from 400 ng input gDNA with an NEBNext Ultra™ II FS DNA Library Prep Kit for Illumina (New England Biolabs) with a total of four

PCR cycles to introduce dual indexed barcodes. Libraries were sequenced in the Max Planck Genome Centre Cologne on a HiSeq2500 system with 2× 250 bp rapid mode using a HiSeq Rapid PE Cluster and Rapid SBS v2 kit. The new sequence data was deposited on GenBank under BioProject PRJNA672693.

Genome size measurements

Novel DNA content measurements for *R. cephalotes* and *R. exaltata* were estimated by flow cytometry. Sample preparation was done according to Loureiro *et al*. (2007). Young leaves of each of the studied plants were chopped simultaneously with their respective reference standard, *Solanum lycopersicum* cv. Stupicke (2C = 1.96 pg, Dolezel *et al*., 1992) for *R. cephalotes* and *Raphanus sativus* L. cv. Saxa (2C = 1.11 pg, Dolezel *et al*., 1992) for *R. exaltata* in a Petri dish (kept on ice) containing 2 mL of Woody Plant Buffer (WPB). The sample was then filtered through a 30-µm disposable mesh filter (CellTrics, SYSMEX, Norderstedt, Germany) with following addition of 50 µg/mL propidium iodide (from a stock of 1 mg/mL; Sigma-Aldrich) and 50 µg/mL RNase (Sigma-Aldrich). Nine replicates per species were made. The samples were measured in a CyFlow Space flow cytometer (SYSMEX) equipped with a green laser (532 nm). Histograms of relative fluorescence were obtained using the software Flomax v.2.3.0. (SYSMEX, Norderstedt, Germany). Mean fluorescence and coefficient of variation were assessed at half of the fluorescence peak. The absolute DNA content (pg/2C) was calculated multiplying the ratio of the G1 peaks by the genome size of the internal standard.

Filtering of TCS reads

As our aim was to characterize the repeat fraction of the sequenced species, we were only interested in the off-target reads, whose abundance is inversely proportional to the efficiency of target sequence enrichment. To get rid of the target reads, we filtered the raw TCS

data by mapping them to a set of 256 sequences representing consensus sequences of the target loci that were enriched in the *Rhynchospora* dataset (Buddenhagen, 2016), saving the unmapped reads for the repeat characterization. Two different mapping algorithms were used for comparison: the *Geneious read mapper* v6.0.3, with custom sensitivity settings (60% Maximum mismatch per read, Index world length = 12, Maximum ambiguity = 8, Kearse *et al.*, 2012) and the BowTie2 v2.4.1 mapper with high sensitivity preset settings (End-to-End alignment, 0 to 800 insert size, report all matches, Langmead & Salzberg, 2012), both implemented in the software Geneious v.7.1.9 (Kearse *et al.*, 2012). These settings were chosen after testing for computational time and mapping results. The Bowtie2 "highest sensitivity" setting presented no major improvement in the mapping results when compared to the faster "high sensitivity" preset. We applied "End-to-End" instead of "Local" alignment to avoid sequence trimming and to better compare to the Geneious Read Mapper results. Additionally to the two mapping algorithms, we uploaded a FASTA file with the 256 target sequences to RepeatExplorer (https://repeatexplorer-elixir.cerit-sc.cz/) prior to the analysis as a Custom Repeat Database (Novak *et al.*, 2013). With this option we could exclude clusters of enriched gene sequences mistakenly identified as repeats from the analysis. In summary, we ended with one GS dataset (comprising of previously published datasets for *R. globosa*, *R. pubera* and *R. tenuis* and the newly sequenced *R. cephalotes* and *R. exaltata* data) and four different "target datasets" for each species: 1) Raw target capture reads; 2) Reads left after mapping with *Geneious read mapper*; 3) Reads left after mapping with BowTie2 Mapper and 4) Reads left after exclusion of "enriched gene clusters" annotated according to a custom Repeat Database (RE).

In silico Repeat Analysis

In order to compare the repeat composition observed in all different datasets, we employed the RepeatExplorer pipeline (Novak *et al.*, 2013). Reads from all datasets of *R. cephalotes, R. exaltata, R. globosa, R. pubera* and *R. tenuis* were uploaded to the platform, filtered by quality with default settings (95% of bases equal to or above the quality cut-off value of 10) and interlaced. Clustering was performed with default settings of 90% similarity over a 55% minimum sequence overlap. The *Find RT Domains* tool and additional database searches (BLASTx) were used to identify protein domains for repeat annotation, and graph layouts of individual clusters were examined interactively using the SeqGrapheR tool (Novak *et al.*, 2013).

Although the entire set of reads from each dataset was uploaded to RepeatExplorer, we used the *Read Sampling* option on the clustering analysis to manually input the number of reads to be analysed, accounting for 0.13× of the genome of each species (**Table 1**). Only for *R. globosa* was this not possible, as no information on genome size was available for this species. In this case, we ran tests with 500,000, 1,000,000 and 2,000,000 reads. However, independent of the number of reads used as input, only 200,061 reads were analysed, always with similar results. Clusters with at least 0.01% genome abundance were automatically annotated and manually checked. We used the TAREAN tool (Novák *et al.*, 2017) available in the RepeatExplorer pipeline to annotate satellite DNAs. Satellites were named based on previous publications (Marques *et al.*, 2015; Ribeiro *et al.*, 2017) or by using the species abbreviation followed by SAT, a number based on the decreasing order of abundance and a hyphen followed by the number of basepairs of the monomer. The consensus monomer sequences of the identified satellite DNAs of each species were compared using DOTTER (Sonnhammer and Durbin, 1995) in order to confirm tandem organization and to identify similarity among repeats from the same family.

Testing Method Performance

To assess if the abundances of the different repeats observed in our raw and filtered target capture datasets were similar, we compared their abundances with the ones observed in the GS datasets. For this, we used the abundance values of the individual repeat families **[Supplementary Table 1]**. As we wanted to check whether the order of abundance of different classes of repetitive elements was similar between the different datasets, we applied Spearman's rank correlation using a t-distribution with *N*-2 degrees of freedom to calculate test significance (*p*-value). The analysis was undertaken with the package *stats* implemented in the software R v. 4.0.2 (R Core Team, 2019). Correlation plots were constructed with the R package ggplot2 (Wickham, 2016).

Repeat amplification, probe labelling and in situ *hybridization*

Primers for the *R. cephalotes* Tyba variant found in the TCS datasets (see results) were designed based on the most conserved region of the consensus sequence (F: 5'-AAGCTATTTGAATGCAATTATGTGC; and R: 5'-AGCGTTTCTAGCCACATTTGA). Genomic DNA (40 ng) of *R. cephalotes* was used for PCR reaction with 1× PCR buffer, 2 mM $MgCl_2$, 0.1 mM of each dNTP, 0.4 μM of each primer, 0.025 U Taq polymerase (Qiagen) and water. The PCR conditions were as follow: 94°C 2 min, 30 cycles of 94°C 50 s, 58°C 50 s and 72°C 1 min and 72°C 10 min. PCR products were labelled with Atto488-dUTP (Jena Bioscience) with a nick translation labelling kit (Jena Bioscience).

Mitotic chromosomes of *R. cephalotes* were prepared from root tips, pre-treated in 2 mM 8-hydroxyquinoline at 10°C for 20 h and fixed in ethanol: acetic acid (3:1 v/v) for 2 h at room temperature and stored at −20°C. Fixed root tips were digested with 2% cellulose, 2% pectinase and 2% pectolyase in citrate buffer (0.01 M sodium citrate and 0.01 M citric acid) for 120 min at 37°C and squashed in a drop of 45% acetic acid. Fluorescent *in situ* hybridization

was performed as described by Aliyeva-Schnorr *et al*. (2015). The hybridization mixture contained 50% (v/v) formamide, 10% (w/v) dextran sulfate, 2×SSC, and 5 ng/µl of the probe. Slides were denatured at 75°C for 5 min, and the final stringency of hybridization was 76%.

Immuno-FISH

To visualize the centromeres of *R. cephalotes*, we performed immunostaining of the centromere-specific histone variant CENH3 with polyclonal antibodies developed for *R. pubera* (RpCENH3, Marques *et al*., 2015). Mitotic preparations were made from root meristems fixed in 4% paraformaldehyde in Tris buffer (10 mM Tris, 10 mM EDTA, 100 mM NaCl, 0.1% Triton, pH 7.5) for 5 minutes on ice under vacuum and for another 25 minutes only on ice. After washing twice in Tris buffer, the roots were chopped in LB01 lysis buffer (15 mM Tris, 2 mM Na_2EDTA, 0.5 mM spermine 4HCl, 80 mM KCl, 20 mM NaCl, 15 mM β-mercaptoethanol, 0.1% Triton X-100, pH 7.5), filtered through a 50 µm filter (CellTrics, Sysmex), diluted 1:10, and subsequently, 100 µl of the diluted suspension were centrifuged onto microscopic slides using a Cytospin3 (Shandon, Germany) as described by Jasencakova *et al*. (2001). Immuno-FISH with anti-RpCENH3 antibodies and the Tyba repeat was performed according to Ishii *et al*. (2015). We used rabbit anti-RpCENH3 (diluted 1:200) as primary antibody and detected it with Cy3-conjugated anti-rabbit IgG (Dianova) secondary antibody (diluted 1:200). Slides were incubated overnight at 4°C and washed three times in 1×PBS before the secondary antibody was applied.

Microscopy

For widefield microscopy, we used an epifluorescence microscope BX61 (Olympus) equipped with a cooled CCD camera (Orca ER, Hamamatsu). To achieve super-resolution of ~120 nm (with a 488 nm laser excitation), we applied spatial structured illumination microscopy

(3D-SIM) using a 63×/1.40 Oil Plan-Apochromat objective of an Elyra PS.1 microscope system and the software ZENBlack from Carl Zeiss GmbH (Weisshart *et al.*, 2016).

Comparative repeat phylogenomics

We employed a repeat abundance-based phylogenetic inference method (see details in Dodsworth *et al.*, 2015) to assess if repeat abundance identified in our TCS reads could be used to resolve phylogenetic relationships, using one of our filtered datasets (BowTie) and the GS dataset for comparison. First, we concatenated reads for our five species with 0.065× coverage, with species-specific codes for each set of reads, and ran a comparative clustering analysis (simultaneous clustering of all species on the dataset) on RepeatExplorer with default settings (Novak *et al.*, 2013). As the sequences were coded with the species names, we could identify the number of reads that each species contributed to each of the generated clusters, which is proportional to the abundance of each repeat in the genome of each species. Parsimony analysis using repeat abundances as quantitative characters was undertaken as described by Dodsworth et al. (2015).

To access the phylogenetic potential of repetitive elements based on sequence similarity, we used the alignment and assembly free (AAF) approach (Fan *et al*., 2015) using all reads identified as repeats by RepeatExplorer in the Bowtie dataset. AAF constructs phylogenies directly from unassembled genome sequence data, bypassing both genome assembly and alignment. Thus, it calculates the statistical properties of the pairwise distances between genomes, allowing it to optimize parameter selection and to perform bootstrapping.

In order to compare our repeat abundance-based phylogeny with a nuclear marker-based phylogeny, we extracted the aligned sequences of 256 loci gathered by target capture (Buddenhagen, 2016) of our five *Rhynchospora* species. For simplification, we used the most general model of DNA substitution GTR + I + G (Abadi *et al.*, 2019). Phylogenetic relationships

were inferred using Bayesian Inference (BI) as implemented on BEAST v.1.8.3 (Drummond and Rambaut, 2007). Two independent runs with four Markov Chain Monte Carlo (MCMC) were conducted, sampling every 1,000 generations for 10,000,000 generations. Each run was evaluated in TRACER v.1.7 (Rambaut *et al.*, 2018) to assess MCMC convergence and a burn-in of 25% was applied. We then obtained the consensus phylogeny and clade posterior probabilities with the "sumt" command.

RESULTS

Efficiency of target sequences filtering

We generated GS data and genome size estimates for *Rhynchospora cephalotes* and *R. exaltata* to add to the already sequenced data of the other three *Rhynchospora* here analysed in order to compare repeat composition from TCS and GS data (**Table 1**). The percentage of filtered reads presented in Table 1 is indicative of the enrichment efficiency, or how much of the final library corresponds to one of the target-genes. The Geneious datasets presented a higher number of filtered reads than the BowTie datasets, with an average of 11.8% of filtered reads for Genious and 8.6% for BowTie (**Table 1**). *R. pubera* presented the smallest number of filtered reads among all five species, with both the BowTie2 and Geneious filters (3.96% and 12.29% respectively). *R. cephalotes* had the highest number of filtered reads among the Geneious datasets, while *R. exaltata* presented the largest proportion among the BowTie datasets. Using the Custom Repeat Database option of RepeatExplorer, the highest amount of "target clusters" was found on *R. exaltata* (18.45%) and the lowest amount was found on *R. tenuis* (6.36%).

Repetitive DNA content of different datasets

To evaluate the quality of repeat characterization from TCS datasets, the genomic proportion of different repetitive element lineages was compared with those observed in the GS datasets (**Fig. 1**). Generally, the proportion of the total repeat fraction observed in the GS datasets was smaller than the ones observed in all the TCS datasets [**Supplementary Figure 1**]. In order to demonstrate whether our filtering strategies were able to improve repeat mining in TCS data, we included the raw TCS datasets in the analysis. As these raw TCS reads contained enriched target sequences, it was expected that these sequences would be present as unclassified clusters by RepeatExplorer. The raw TCS datasets presented the highest values for total repetitive fraction in almost all species (including the unclassified, putative target sequences), with the exception of *R. globosa,* in which the Geneious dataset showed a total of 49.41% repeat proportion against 46.74% on the raw dataset. This is also reflected in the difference of number of clusters representing at least 0.01% of total genomic proportion formed in the clustering analysis. While the GS datasets presented cluster numbers varying from 155 to 294, filtered TCS datasets ranged from 462 to 589 clusters and raw TCS datasets ranged from 628 to 751. These differences are mostly due to the discrepancy in the proportion of unclassified repetitive elements in the different datasets [**Supplementary Figure 1**]. Unclassified repeat proportion on GS varied from 6.21% in *R. exaltata* to 14.70% in *R. globosa,* while in TCS it accounted for to up to 43.63% (raw TCS of *R. exaltata*). Overall, proportion of unclassified repeats was smaller in all filtered datasets when compared to raw TCS (**Fig. 2a**). Additional mapping to the original 256 target regions and separated BLAST searches with conserved domains did not produce any matches for the unclassified clusters. In contrast, BLASTx of highly abundant unclassified clusters of the TCS datasets showed similarity to coding sequences for proteins such as glycosylphosphatidylinositol anchor protein and oligomeric golgi complex subunits. This confirmed that at least part of the excess of unclassified clusters in the TCS data was a byproduct of accidentally enriched non-repetitive genomic regions.

Furthermore, there was a high number of repetitive element lineages that were not found in the GS dataset but were found in TCS datasets **[Supplementary Table 1**]. These additional repeats could be low-abundance elements that, although not abundant enough to be detected in the GS datasets, could have been accidentally enriched in the TCS datasets.

The filtering strategies did not largely impact satellite DNA abundance in the analysed species, with the exception of *R. exaltata*, in which it was possible to identify ~4× more satellite reads in the BowTie2 and Geneious datasets than in the raw datasets. Also in *R. exaltata*, there was a huge discrepancyin the abundance of satellites, much higher in the GS dataset when compared to all TCS datasets (**Fig. 1**). However, the two satellite DNAs responsible for this abundance difference were also found in all TCS datasets [**Supplementary Table 2, Supplementary Figure 2**], although in smaller proportions. The amount of satellite DNA was generally lower in the TCS datasets, with the exception of *R. tenuis*, in which all TCS datasets presented a small increase in satellite abundance when compared to GS (**Fig. 1**). In this case, clusters formed by unfiltered enriched sequences could have masked the identification of low-abundance repeats. Although in some species some of the satellites found in GS data were not present in every dataset, the most abundant for each species could be identified in all of the TCS datasets. Similarly, some satellites found in TCS datasets were not found in the GS datasets, possibly being low-abundance satellites accidentally enriched [**Supplementary Table 2, Supplementary Figure 2**].

For mobile elements, there was a general agreement in the order of abundance of repeat types found in the GS and TCS datasets, with filtered datasets showing increase in the proportion of annotated elements when compared to the raw TCS dataset (**Fig. 1**), with a few exceptions. For example, in *R. exaltata*, LTRs from the Ty3/gypsy superfamily were more abundant in the raw TCS, BowTie2 and RE datasets than in the GS datasets (**Fig. 1**). We also compared the abundances at lineage level **[Supplementary Table 1]**. Patterns of abundance of

LTR families from GS and TCS datasets were similar, with most of the genomic abundance of Ty1/copia and Ty3/gypsy superfamilies being result of the amplification of up to four main lineages **[Supplementary Table 1]**. Generally, LTRs found in GS with abundance as low as 0.01% could also be identified in the target datasets. Surprisingly, in all five species, a greater diversity of LTR retroelements was observed in the target capture datasets when compared to GS **[Supplementary Table 1]**. This led to a few interesting discrepancies, such as in *R. pubera,* where target capture datasets showed high abundance of Ty3/gypsy/Retand (1 to 1.3%) and Ty3/gypsy/Tekay elements (0.46 to 0.50%), despite those not being found in GS data.

To statistically compare the results obtained in the different datasets, we checked for a correlation between the classified repeat abundances observed (at lineage level, when possible) in all target capture datasets with the ones observed in the GS datasets (**Fig. 3**). Although all tests showed significant correlations ($p < 0.05$), the strength of the correlation varied depending on the dataset. Raw TCS datasets had the weakest correlation in all five species, with filtering of the targeted sequences generally improving the correlation with the GS dataset (**Fig. 2b**). The strongest correlations for *R. globosa* ($r = 0.92$), *R. pubera* ($r = 0.74$) and *R. tenuis* ($r = 0.79$) were observed with the Geneious dataset, while for *R. cephalotes* and *R. exaltata* the best correlation was observed on the BowTie dataset ($r = 0.78$ and 0.75). The RE filtering for *R. globosa* and *R. tenuis* did not improve the correlation when compared to the raw TCS dataset ($r = 0.85$ and $r = 0.66$ respectively). However, it also showed as strong a correlation as BowTie for *R. cephalotes* ($r = 0.78$) and as Geneious for *R. pubera* ($r = 0.74$). In addition to this, we checked if there was a significant correlation between enrichment efficiency (the proportion of the genomic library that hybridized to a target probe) and the repeat characterization efficiency (the proportion of classified repeats given by the RE analysis). Even though our sampling was small (n = 5), we could see a clear trend of inverse correlation between these values ($p = 0.02$,

R^2 = 0.83), which shows that highly efficient library enrichment may impair repeat characterization in off-target reads.

Chromosomal localization of the satellite DNA found in the TCS dataset

To test whether it was possible to use the TCS data to investigate the chromosomal repeat distribution by FISH, we chose the most abundant repetitive element found in the *R. cephalotes* TCS datasets. This repeat was a 172-bp satellite DNA with 60% sequence similarity to Tyba of *R. pubera* (Marques *et al.*, 2015). The *in situ* hybridization pattern of the *R. cephalotes* Tyba (RcTyba) variant was similar to the distribution reported in *R. pubera*. Small foci appeared in interphase nuclei, and a continuous line along both condensed chromatids occurred at all pro- and metaphase chromosomes. Via immuno-FISH using a CENH3-specific antibody and RcTyba repeats, respectively, the holocentric centromere structure of *R. cephalotes* has been confirmed due to the co-localization of CENH3 and RcTyba (**Fig. 4**).

Repeat abundance and structure found in TCS data reflect phylogenetic relationships

We used repeat abundances obtained by comparative clustering analysis of BowTie datasets of our five *Rhynchospora* species to reconstruct phylogenetic relationships. In the comparative clustering analysis of the GS dataset, 1,754,326 concatenated reads were analysed, forming 450 clusters with at least 0.01% genomic abundance. For the BowTie dataset, 1,186,605 of concatenated reads were analysed, with 582 clusters representing at least 0.01% of total genomic abundance. Repeat composition varied among species, with the largest clusters of each species being almost or completely absent in the others (**Fig. 5a**). By using the first 150 most abundant clusters of the comparative analysis, we were able to reconstruct the phylogenetic relationships among the five *Rhynchospora* species for both the BowTie (**Fig. 5b**)

and GS datasets (**Fig. 5c**) with high bootstrap support (mean BS = 100 and 98.3 respectively). Using the reads from all clusters identified in the BowTie dataset comparative analysis, the AAF analysis yielded the same relationships with high bootstrap support (mean BS = 100, **Fig. 4d**). Branch lengths of the abundance-based analysis were significantly higher than for the AAF analysis. Despite this, the species relationships observed in the repeat-based phylogenies were congruent with the ones retrieved in the Bayesian analysis with 256 concatenated target loci, with *R. cephalotes* + *R. exaltata* forming a clade sister to *R. pubera* + *R. tenuis*, and *R. globosa* sister to both clades (**Fig. 5e**).

DISCUSSION

TCS data can be used to identify highly abundant repeats

We were able to find most of the repeat diversity of five *Rhynchospora* species using filtered and unfiltered TCS reads. Depending on the filtering strategy employed, the *Rhynchospora* data used here showed low abundance of on-target reads, indicating a considerable proportion of off-target reads suitable for repeat analysis. Target-based sequencing approaches often have varied enrichment efficiency, with on-target enriched reads representing as low as 5% of the final genomic library (Johnson *et al.*, 2019). This is particularly the case with universal kits that are designed to work across large taxonomic groups, at the cost of enrichment efficiency. Thus, it is believed that the off-target sequence reads from such libraries can be used in a similar fashion to genome skimming sequencing, an approach often described as Hyb-Seq (Weitemier *et al.*, 2014; Dodsworth *et al.*, 2019). This approach has been used for plastome assembly in several species (Weitemier *et al.*, 2014; Schmickl *et al.*, 2016), as well as for ribosomal DNA profile comparisons, in which GS and TCS data showed satisfactory correlation (Sproul *et al.*, 2020). However, this method was yet to be tested for identification of high copy repeats such as satellite DNA and transposable elements.

Our results show that there is a moderate rank correlation between the abundances of annotated repeats obtained by analysing genome skimming and raw target capture datasets, and that this correlation increases when analysing off-target reads only. The significant correlation indexes showed that although annotated repeat proportions of our filtered target datasets are not identical to those observed in GS, high copy repeats can be sufficiently identified in off-target reads. Although the Geneious dataset presented the highest proportion of filtered reads and highest correlation index with GS data for three of the five species, the BowTie2 and RE approaches were also sufficient for increasing the correlation index in most cases. Furthermore, different settings for the mappers, such as the "Local" Read Alignment setting for BowTie2, can be tested in order to improve mapping to target sequences. Therefore, any of the filtering strategies presented here can produce sufficiently accurate results in regards to repeat identification and order of abundance when compared with the traditional GS approach.

Although the total abundance of satellite DNAs varied between GS and TCS datasets, we were able to identify the most abundant satellite DNA families of all five *Rhynchospora* species in all target datasets. While some low-abundance satellites (<0.2% of genomic abundance) found in GS data were not found in the target datasets, others with abundances as low as 0.09% were found in all datasets, suggesting that satellite abundance does not necessarily affect its presence in the off-target reads **[Supplementary Table 2]**. More importantly, various satellites found in the TCS datasets were not found in the GS dataset **[Supplementary Table 2]**, probably being low-abundant satellites accidentally enriched in the sequencing process. Low-abundant satellites may sometimes not be detected by RepeatExplorer, requiring additional filtering to be detected (Ruiz-Ruano *et al.*, 2016), which could explain the absence of these satellites in our GS datasets.

One of the benefits of using *Rhynchospora* as a model for this study was the fact that it possesses satellite DNAs with varying chromosomal distributions. Tyba clusters are dispersed,

forming a linear distribution along the metaphase holocentromeres of *R. pubera*, *R. ciliata*, *R. cephalotes* and *R. tenuis* (Marques *et al*., 2015; Ribeiro *et al*., 2017; this study). However, other satellites in *Rhynchospora*, such as *Rg*SAT1-186 in *R. globosa* form localized blocks (Ribeiro *et al*., 2017). Off-target sequences used in Hyb-Seq are often adjacent to the targeted gene regions (Weitemier *et al.*, 2014; Dodsworth *et al.*, 2019), rasing the possibility that our analysis would preferentially identify widespread repeats such as Tyba, which are interspersed with genic regions (Marques *et al.*, 2015). However, *Rg*SAT1-186, identified as sub-terminal clusters in the chromosomes of *R. globosa* (Ribeiro *et al.*, 2017), was identified as the most abundant cluster on all the *R. globosa* TCS datasets, showing that the chromosomal distribution of a satellite DNA was not interfering in the randomness of the off-target reads [**Supplementary Figure 1**].

As well as confirming its presence in the target datasets of *R. pubera* and *R. tenuis*, we were able to find novel Tyba variants in *R. cephalotes* and in *R. exaltata*, with 60% and 57% sequence similarity to *Rp*Tyba, respectively. *R. globosa* did not present any Tyba variant, in concordance with the results of Ribeiro *et al*. (2017). The abundance of Tyba in *R. pubera* was very low in the target datasets (~0.14%) compared to the GS results (~2.8% here, 3.6% in Marques *et al*., 2015). In our *R. pubera* TCS datasets, the most abundant tandem repeat was *Rp*SAT5-287, which appeared as only the fifth most abundant in the GS dataset **[Supplementary Table 1]**. This satellite was not found in the previous *R. pubera* characterization and may indicate the potential to discover additional low-abundant sequences using TCS datasets. Although fast rates of evolution for satellite DNAs may lead to intraspecific abundance variation (Ceccarelli *et al.*, 2011), this 20-fold difference is probably too high to be a product of differences between the individuals used for each sequencing method. Satellite abundances estimated by TAREAN depend on several factors, such as sequence coverage, monomer homogeneity and similarity with other repeats (Novák *et al.*, 2017). As abundance

information gathered by short reads can grossly underestimate the true abundance of this repeat type (i.e. Ribeiro *et al.*, 2020), caution is needed when interpreting TCS-yielded genomic abundances, though similar caution is needed with GS data as well.

Transposable elements, particularly LTR-retrotransposons, are often the largest fraction of repetitive DNA in plants (Galindo-González *et al.*, 2017). In our GS datasets, LTR−Ty1/copia was the most abundant repeat type in three out of five species. The TCS datasets yielded similar results to GS, with a few discrepancies, especially in *R. exaltata* (predominance of Ty3/gypsy instead of Ty1/copia) and *R. pubera* (significantly higher proportion of Ty3/gypsy than in GS). As discussed for the satellites, this discrepancies could be the result of some repeat lineages being accidentally enriched during the TCS procedure, which should demand caution when interpreting repeat abundances in these type of data. Additionally, the intraspecific variability of repeat abundance could account for some of the variation observed here, since GS and TCS datasets came from different individuals. Intraspecific repeat variation ranges from low (Renny-Byfield and Baumgarten, 2020) to high among natural populations and even within organism (Shams and Raskina, 2018). Thus, it is possible that part of the discrepancies could also be caused by natural differences in the genomic abundance of some repetitive lineages within species. In order to do a more detailed comparison, we used the individual lineage abundance values for our correlation analysis. The high correlation rates indicate that different LTR lineages, although varying in abundance, contributed similarly to the repetitive fraction in GS and filtered target datasets. Our results show that in addition to being able to find the majority of amplified lineages of LTR retrotransposons, we could also find lineages with abundances as low as 0.01% in the target datasets. We also find a higher diversity of LTR lineages in the target datasets than in the GS data, with a few of these "extra" lineages being over-abundant when compared to the GS dataset **[Supplementary Table 1]**. These lineages were absent in the GS results probably due to

masking by the highly abundant satellite DNA clusters, which were underestimated in some of our TCS datasets. Nonetheless, the fact that we could identify even low abundance LTRs in the target datasets, coupled with the moderate correlation with the GS-yielded abundances, indicate that off-target reads from target sequencing may be sufficient to identify most of the LTR retrotransposons in a genome.

It is important to note that, although the patterns of repeat abundance are highly similar, the numerical difference observed for some repeat classes shows that combining GS and TCS data may produce inconsistent results. Although it is possible to mine off-target reads for predominant repeats, the exact abundance values are not as accurate as those yielded by GS data, which is still the most recommended way of gather such information (Dodsworth, 2015). Another important point is that technical bias could still be one big factor in the reliability of TCS data for repeat mining. With our limited sample, it was already possible to find a high inverse correlation between the proportion of classified repeats and enrichment efficiency (or the proportion of on-target reads in the final TCS library). Enrichment efficiency of a single target-capture kit can vary greatly among distant species (i.e.: 5% to 68% variation in the species tested by Johnson *et al.*, 2019). Future uses of off-target TCS reads may have to take into account that species with highly efficient enrichment may not produce reliable repeat characterization. As good capture efficiency (number of target genes in the final library) can be achieved with low enrichment efficiency (Dodsworth, 2015; Johnson *et al.*, 2019), one possible strategy for future projects would be to sequence a portion of the unriched library, in order to maximize the usefulness of the final dataset.

TCS data can be used to develop cytogenetic probes

We tested whether we could use repeat information obtained from TCS data to develop probes for cytogenetic techniques such as fluorescent *in situ* hybridization (FISH). Highly abundant

repetitive elements are frequently chosen for FISH experiments, as they are easier to visualize on chromosomes than low abundance repeats and can often be important components of predominantly heterochromatic and centromeric regions (Marques *et al.*, 2015 Bilinski *et al.*, 2017; Ávila Robledillo *et al.*, 2018). In our *Rhynchospora* species, the most abundant cluster in all datasets was a satellite DNA. For *R. tenuis* and *R. globosa* we found the same satellites found previously by Ribeiro *et al*. (2017) as the most abundant satellites (Tyba and *Rg*SAT1-186 respectively).

Similar to previous results on other *Rhynchospora* (Marques *et al.*, 2015; Ribeiro *et al.*, 2017), the Tyba variant from *R. cephalotes* presented a dispersed distribution in interphase nuclei and a line-like pattern along the sister chromatids of condensed chromosomes. Co-localization with CENH3 further confirmed the holocentromere specific localization of Tyba. The conservation of the holocentromeric distribution of *Rc*Tyba on *R. cephalotes* strengthens its putative role in centromere function as proposed previously (Marques *et al.*, 2015; Ribeiro *et al.*, 2017). It also points to a remarkably old origin for this satellite and its holocentromeric association, sharing a common ancestor between 35-46 My (95% credible interval; Buddenhagen, 2016). A large-scale survey of Tyba in the entire *Rhynchospora* genus could help to further elucidate the evolution of this satellite and its association with the holocentromere. Although GS is still the most reliable way to gather repeat information, it is still expensive to sequence a great number of species. With the growing availability of TCS data (Johnson *et al.*, 2019; Andermann *et al.*, 2020), the possibility of recycling off-target reads to mine highly abundant repeats demonstrated here can be a cost-efficient alternative for large-scale cytogenomic investigations.

TCS data can be used to construct repeat-based phylogenies

In order to test if repeat abundances observed in target datasets are accurate enough to reconstruct phylogenetic relationships, we applied the methodology described by Dodsworth *et al.* (2015) to our BowTie and GS datasets as well as using an assembly and alignment free method (Fan *et al.*, 2015) using the total set of reads output by the comparative repeat analysis on the BowTie dataset. Dodsworth et al.'s approach takes into account the assumption that, as repeat abundance changes primarily through random genetic drift (Jurka *et al.*, 2011), they can be used as selection-free characters for phylogenetic reconstruction. On the other hand, AAF methods have been shown to identify potentially useful markers for taxonomic resolution from genome skimming datasets (Bohmann *et al.*, 2020). Both analyses were successful in reconstructing the major relationships between the five *Rhynchospora* species with maximal bootstrap support. Our results not only corroborate that AAF methods can be useful for repetitive sequence data, but also show that it can be employed on the off-target portion of TCS data. These sequence similarity-based approaches (e.g. Vitales *et al.*, 2020) may be more appropriate for groups where no genome skimming data is available for verifying the accuracy of repeat abundance in the target dataset, when abundance-based approaches provide less resolved trees or when genome sizes are unknown.

The phylogenetic relationships retrieved by the repeat abundance and AAF-based phylogeny were not only congruent with the GS based analysis and a Bayesian tree of the 256 target regions, but also with recent studies in the genus based on other phylogenetic data (Buddenhagen *et al.*, 2016; Ribeiro *et al.*, 2018). The fine resolution in both repeat abundance analysis is mainly due to the significant intra-specific difference in repeat composition observed in the BowTie dataset and corroborated by the GS analysis. Although it is common to closely related species to share similar repeat profiles, *Rhynchospora* is a fairly old genus and some of the species here analysed diverge by several My (Buddenhagen, 2016). Our results show the potential of repeats from off-target reads to be used as an additional phylogenetically

informative dataset, from a completely different part of the genome typically used for phylogenetic studies. Target capture-based sequencing already offers the opportunity to construct robust phylogenies with hundreds of informative markers and also to assemble whole plastomes and other organellar DNAs via the off-target reads to infer phylogenetic relationships, and nuclear-organellar discordance (Dodsworth *et al.*, 2019). Repeat-based phylogenies offer an additional strategy, based on the same TCS datasets, potentially uncovering nuclear intragenomic (in)congruence, while further increasing the usefulness of TCS datasets. Even in cases were repeat-based phylogenies cannot offer additional species evolution insights, it can help to understand the evolution of certain repetitive sequences across a group. The robustness of the repetitive DNA information obtained from our target datasets can prove useful in a variety of phylogenomic approaches, such as similarity-based repeat phylogenies (Vitales *et al.*, 2020), as well as the alignment-free and abundance-based methods presented here.

ACKNOWLEGMENTS

The authors are grateful to Dr. Magdalena Vaio (Facultad de Agronomía, Uruguay) for providing comments and suggestions for the manuscript and to Msc. Erton Almeida for the collection of *Rhynchospora cephalotes*.

FUNDING

This study was supported in part by the Coordenacão de Aperfeicoamento de Pessoal de Nıvel Superior–Brasil (CAPES) [Finance Code 001], CAPES-PRINT [project number 88887.363884/2019-00 (LC)], and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnologico) [grant number 141037/2018-0 (LC)].

LITERATURE CITED

Abadi S, Azouri D, Pupko T, Mayrose I. **2019**. Model selection may not be a mandatory step for phylogeny reconstruction. *Nature Communications* **10**: 934. doi: 10.1038/s41467-019-08822-w.

Albert TJ, Molla MN, Muzny DM, *et al.* 2007. Direct selection of human genomic loci by microarray hybridization. *Nature Methods* **4**: 903–905. doi: 10.1038/nmeth1111.

Aliyeva-Schnorr L, Beier S, Karafiátová M, *et al.* 2015. Cytogenetic mapping with centromeric bacterial artificial chromosomes contigs shows that this recombination-poor region comprises more than half of barley chromosome 3H. *The Plant Journal* **84**: 385–394. doi: 10.1111/tpj.13006.

Andermann T, Torres Jiménez MF, Matos-Maraví P, *et al.* 2020. A Guide to Carrying Out a Phylogenomic Target Sequence Capture Project. *Frontiers in Genetics* **10**: 1407. doi: 10.3389/fgene.2019.01407.

Ávila Robledillo L, Koblížková A, Novák P, *et al.* 2018. Satellite DNA in Vicia faba is characterized by remarkable diversity in its sequence composition, association with centromeres, and replication timing. *Scientific Reports* **8**. doi: 10.1038/s41598-018-24196-3.

Bilinski P, Albert PS, Berg JJ, *et al.* 2017. Parallel Altitudinal Clines Reveal Adaptive Evolution Of Genome Size In *Zea mays*. doi: 10.1371/journal.pgen.1007162.

Bohmann K, Mirarab S, Bafna V, Gilbert MTP. **2020**. Beyond DNA barcoding: The unrealized potential of genome skim data in sample identification. *Molecular Ecology* **29**: 2521–2534. doi: 10.1111/mec.15507.

Bolsheva NL, Melnikova NV, Kirov IV, *et al.* 2019. Characterization of repeated DNA sequences in genomes of blue-flowered flax. *BMC Evolutionary Biology* **19**: 49. doi: 10.1186/s12862-019-1375-6.

Buddenhagen CE. **2016**. A view of Rhynchosporeae (Cyperaceae) diversification before and after the application of anchored phylogenomics across the angiosperms. PhD Thesis, Florida State University, USA

Buddenhagen C, Lemmon AR, Lemmon EM, *et al.* 2016. *Anchored Phylogenomics of Angiosperms I: Assessing the Robustness of Phylogenetic Estimates*. Evolutionary Biology. doi: 10.1101/086298

Bureš P, Zedek F, Markova M. **2013**. Holocentric Chromosomes In: *Plant Genome Diversity Volume 2*. Vienna: Springer Vienna, 187–204.

Ceccarelli M, Sarri V, Caceres ME, Cionini PG. **2011**. Intraspecific genotypic diversity in plants (P Donini, Ed.). *Genome* **54**: 701–709. doi: 10.1139/g11-039.

Cheng Z, Dong F, Langdon T, *et al.* 2002. Functional Rice Centromeres Are Marked by a Satellite Repeat and a Centromere-Specific Retrotransposon. *The Plant Cell* **14**: 1691–1704. doi: 10.1105/tpc.003079

Čížková J, Hřibová E, Humplíková L, Christelová P, Suchánková P, Doležel J. **2013**. Molecular Analysis and Genomic Organization of Major DNA Satellites in Banana (Musa spp.) (K Kashkush, Ed.). *PLoS ONE* **8**: e54808. doi: 10.1371/journal.pone.0054808.

Cosart T, Beja-Pereira A, Chen S, Ng SB, Shendure J, Luikart G. **2011**. Exome-wide DNA capture and next generation sequencing in domestic and wild species. *BMC Genomics* **12**: 347. doi: 10.1186/1471-2164-12-347.

Dodsworth S. **2015**. Genome skimming for next-generation biodiversity analysis. *Trends in Plant Science* **20**: 525–527. doi: 10.1016/j.tplants.2015.06.012.

Dodsworth S, Chase MW, Kelly LJ, *et al.* 2015. Genomic Repeat Abundances Contain Phylogenetic Signal. *Systematic Biology* **64**: 112–126. doi: 10.1093/sysbio/syu080.

Dodsworth S, Jang T-S, Struebig M, Chase MW, Weiss-Schneeweiss H, Leitch AR. **2017**. Genome-wide repeat dynamics reflect phylogenetic distance in closely related allotetraploid *Nicotiana* (Solanaceae). *Plant Systematics and Evolution* **303**: 1013–1020. doi: 10.1007/s00606-016-1356-9.

Dodsworth S, Pokorny L, Johnson MG, *et al.* 2019. Hyb-Seq for Flowering Plant Systematics. *Trends in Plant Science* **24**: 887–891. doi: 10.1016/j.tplants.2019.07.011.

Dolezel J, Sgorbati S, Lucretti S. **1992**. Comparison of three DNA fluorochromes for flow cytometric estimation of nuclear DNA content in plants. *Physiologia Plantarum* **85**: 625–631. doi: 10.1111/j.1399-3054.1992.tb04764.x.

Drummond AJ, Rambaut A. **2007**. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7**: 214. doi: 10.1186/1471-2148-7-214.

Eaton DAR, Spriggs EL, Park B, Donoghue MJ. **2016**. Misconceptions on Missing Data in RAD-seq Phylogenetics with a Deep-scale Example from Flowering Plants. *Systematic Biology*: syw092. doi: 10.1093/sysbio/syw092.

Elliott TA, Gregory TR. **2015**. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**: 20140331. doi: 10.1098/rstb.2014.0331.

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. **2012**. Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales. *Systematic Biology* **61**: 717–726. doi: 10.1093/sysbio/sys004.

Fan H, Ives AR, Surget-Groba Y, Cannon CH. **2015**. An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. *BMC Genomics* **16**: 522. doi: 10.1186/s12864-015-1647-5.

Galindo-González L, Mhiri C, Deyholos MK, Grandbastien M-A. **2017**. LTR-retrotransposons in plants: Engines of evolution. *Gene* **626**: 14–25. doi: 10.1016/j.gene.2017.04.051.

Gnirke A, Melnikov A, Maguire J, *et al.* 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology* **27**: 182–189. doi: 10.1038/nbt.1523.

Guignard MS, Nichols RA, Knell RJ, *et al.* 2016. Genome size and ploidy influence angiosperm species' biomass under nitrogen and phosphorus limitation. *New Phytologist* **210**: 1195–1206. doi: 10.1111/nph.13881.

Heyduk K, Trapnell DW, Barrett CF, Leebens-Mack J. **2016**. Phylogenomic analyses of species relationships in the genus *Sabal* (Arecaceae) using targeted sequence capture. *Biological Journal of the Linnean Society* **117**: 106–120. doi: 10.1111/bij.12551.

Houben A, Schubert I. **2003**. DNA and proteins of plant centromeres. *Current Opinion in Plant Biology* **6**: 554–560. doi: 10.1016/j.pbi.2003.09.007.

Ilves KL, López-Fernández H. **2014**. A targeted next-generation sequencing toolkit for exon-based cichlid phylogenomics. *Molecular Ecology Resources* **14**: 802–811. doi: 10.1111/1755-0998.12222.

Ishii T, Sunamura N, Matsumoto A, Eltayeb AE, Tsujimoto H. **2015**. Preferential recruitment of the maternal centromere-specific histone H3 (CENH3) in oat (*Avena sativa* L.) × pearl millet (*Pennisetum glaucum* L.) hybrid embryos. *Chromosome Research* **23**: 709–718. doi: 10.1007/s10577-015-9477-5

Jasencakova Z, Meister A, Schubert I. **2001**. Chromatin organization and its relation to replication and histone acetylation during the cell cycle in barley. *Chromosoma* **110**: 83–92. doi: 10.1007/s004120100132.

Johnson MG, Pokorny L, Dodsworth S, *et al.* 2019. A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering (S Renner, Ed.). *Systematic Biology* **68**: 594–606. doi: 10.1093/sysbio/syy086.

Jurka J, Bao W, Kojima KK. **2011**. Families of transposable elements, population structure and the origin of species. *Biology Direct* **6**: 44. doi: 10.1186/1745-6150-6-44.

Kearse M, Moir R, Wilson A, *et al.* 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**: 1647–1649. doi: 10.1093/bioinformatics/bts199.

Koo D-H, Hong CP, Batley J, *et al.* 2011. Rapid divergence of repetitive DNAs in Brassica relatives. *Genomics* **97**: 173–185. doi: 10.1016/j.ygeno.2010.12.002.

Langmead B, Salzberg SL. **2012**. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357–359. doi: 10.1038/nmeth.1923.

Lemmon AR, Emme SA, Lemmon EM. **2012**. Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics. *Systematic Biology* **61**: 727–744. doi: 10.1093/sysbio/sys049.

Loureiro J, Rodriguez E, Dolezel J, Santos C. **2007**. Two new nuclear isolation buffers for plant DNA flow cytometry: a test with 37 species. *Annals of Botany* **100**: 875–888. doi: 10.1093/aob/mcm152.

Lyu H, He Z, Wu C-I, Shi S. **2018**. Convergent adaptive evolution in marginal environments: unloading transposable elements as a common strategy among mangrove genomes. *New Phytologist* **217**: 428–438. doi: 10.1111/nph.14784.

Macas J, Novák P, Pellicer J, *et al.* 2015. In Depth Characterization of Repetitive DNA in 23 Plant Genomes Reveals Sources of Genome Size Variation in the Legume Tribe Fabeae (A Houben, Ed.). *PLOS ONE* **10**: e0143424. doi: 10.1371/journal.pone.0143424.

Mandel JR, Dikow RB, Funk VA, *et al.* 2014. A Target Enrichment Method for Gathering Phylogenetic Information from Hundreds of Loci: An Example from the Compositae. *Applications in Plant Sciences* **2**: 1300085. doi: 10.3732/apps.1300085.

Marques A, Ribeiro T, Neumann P, *et al.* 2015. Holocentromeres in *Rhynchospora* are associated with genome-wide centromere-specific repeat arrays interspersed among euchromatin. *Proceedings of the National Academy of Sciences* **112**: 13633–13638. doi: 10.1073/pnas.1512255112.

Martín-Peciña M, Ruiz-Ruano FJ, Camacho JPM, Dodsworth S. **2019**. Phylogenetic signal of genomic repeat abundances can be distorted by random homoplasy: a case study from hominid primates. *Zoological Journal of the Linnean Society* **185**: 543–554. doi: 10.1093/zoolinnean/zly077.

Mascagni F, Vangelisti A, Giordani T, Cavallini A, Natali L. **2020**. A computational comparative study of the repetitive DNA in the genus *Quercus* L. *Tree Genetics & Genomes* **16**: 11. doi: 10.1007/s11295-019-1401-2.

Nagaki K, Talbert PB, Zhong CX, Dawe RK, Henikoff S, Jiang J. **2003**. Chromatin immunoprecipitation reveals that the 180-bp satellite repeat is the key functional DNA element of Arabidopsis thaliana centromeres. *Genetics* **163**: 1221–1225.

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**. doi: 10.1186/s13100-018-0144-1.

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 Research* **45**: e111–e111. doi: 10.1093/nar/gkx257.

Novak 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**: 792–793. doi: 10.1093/bioinformatics/btt054.

Pellicer J, Fay MF, Leitch IJ. **2010**. The largest eukaryotic genome of them all?: THE LARGEST EUKARYOTIC GENOME? *Botanical Journal of the Linnean Society* **164**: 10–15. doi: 10.1111/j.1095-8339.2010.01072.x.

R Core Team. **2019**. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. **2018**. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7 (E Susko, Ed.). *Systematic Biology* **67**: 901–904. doi: 10.1093/sysbio/syy032.

Renny-Byfield S, Baumgarten A. 2020. Repetitive DNA content in the maize genome is uncoupled from population stratification at SNP loci. *BMC Genomics* **21**: 98. doi: 10.1186/s12864-020-6517-0

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**: 263–272. doi: 10.1007/s00709-017-1154-4.

Ribeiro T, Marques A, Novák P, *et al.* 2017. Centromeric and non-centromeric satellite DNA organisation differs in holocentric *Rhynchospora* species. *Chromosoma* **126**: 325–335. doi: 10.1007/s00412-016-0616-3.

Ribeiro T, Vasconcelos E, dos Santos KGB, Vaio M, Brasileiro-Vidal AC, Pedrosa-Harand A. **2020**. Diversity of repetitive sequences within compact genomes of *Phaseolus* L. beans and allied genera *Cajanus* L. and *Vigna* Savi. *Chromosome Research* **28**: 139–153. doi: 10.1007/s10577-019-09618-w.

Ruiz-Ruano FJ, López-León MD, Cabrero J, Camacho JPM. **2016**. High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Scientific Reports* **6**. doi: 10.1038/srep28333.

Sarmashghi S, Bohmann K, P. Gilbert MT, Bafna V, Mirarab S. **2019**. Skmer: assembly-free and alignment-free sample identification using genome skims. *Genome Biology* **20**: 34. doi: 10.1186/s13059-019-1632-4.

Sass C, Iles WJD, Barrett CF, Smith SY, Specht CD. **2016**. Revisiting the Zingiberales: using multiplexed exon capture to resolve ancient and recent phylogenetic splits in a charismatic plant lineage. *PeerJ* **4**: e1584. doi: 10.7717/peerj.1584.

Schmickl R, Liston A, Zeisek V, *et al.* 2016. Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African *Oxalis* (Oxalidaceae). *Molecular Ecology Resources* **16**: 1124–1135. doi: 10.1111/1755-0998.12487.

Schrader L, Schmitz J. **2019**. The impact of transposable elements in adaptive evolution. *Molecular Ecology* **28**: 1537–1549. doi: 10.1111/mec.14794.

Shams I, Raskina O. 2018. Intraspecific and intraorganismal copy number dynamics of retrotransposons and tandem repeat in *Aegilops speltoides* Tausch (Poaceae, Triticeae). *Protoplasma* **255:** 1023–1038 doi: 10.1007/s00709-018-1212-6

Sonnhammer ELL, Durbin R. **1995**. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167**: GC1–GC10. doi: 10.1016/0378-1119(95)00714-8.

Souza G, Costa L, Guignard MS, *et al.* 2019. 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* **38**: 13–23. doi: 10.1016/j.ppees.2019.03.002.

Sproul JS, Barton LM, Maddison DR. 2020. Repetitive DNA Profiles Reveal Evidence of Rapid Genome Evolution and Reflect Species Boundaries in Ground Beetles. *Systematic Biology* **69**: 1137-1148. doi: 10.1093/sysbio/syaa030

Straub SCK, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A. **2012**. Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany* **99**: 349–364. doi: 10.3732/ajb.1100335.

Thomas WmW, Araújo AC, Alves MV. **2009**. A Preliminary Molecular Phylogeny of the Rhynchosporeae (Cyperaceae). *The Botanical Review* **75**: 22–29. doi: 10.1007/s12229-008-9023-7.

Van-Lume B, Esposito T, Diniz-Filho JAF, Gagnon E, Lewis GP, Souza G. **2017**. Heterochromatic and cytomolecular diversification in the Caesalpinia group (Leguminosae): Relationships between phylogenetic and cytogeographical data. *Perspectives in Plant Ecology, Evolution and Systematics* **29**: 51–63. doi: 10.1016/j.ppees.2017.11.004.

Vitales D, Garcia S, Dodsworth S. **2020**. Reconstructing Phylogenetic Relationships Based on Repeat Sequence Similarities. *Molecular Phylogenetics and Evolution* **147**: 106766. doi: 10.1016/j.ympev.2020.106766.

Wang H-J, Li W-T, Liu Y-N, Yang F-S, Wang X-Q. **2017**. Resolving interspecific relationships within evolutionarily young lineages using RNA-seq data: An example from Pedicularis section Cyathophora (Orobanchaceae). *Molecular Phylogenetics and Evolution* **107**: 345–355. doi: 10.1016/j.ympev.2016.11.018.

Weisshart K, Fuchs J, Schubert V. **2016**. Structured Illumination Microscopy (SIM) and Photoactivated Localization Microscopy (PALM) to Analyze the Abundance and Distribution of RNA Polymerase II Molecules on Flow-sorted Arabidopsis Nuclei. *BIO-PROTOCOL* **6**. doi: 10.21769/BioProtoc.1725.

Weiss-Schneeweiss H, Leitch AR, McCann J, Macas J. **2015**. Exploring the repeats' landscape and its impact on genome evolution and plant diversification In: Germany: Koeltz Scientific Books, .

Weitemier K, Straub SCK, Cronn RC, *et al.* 2014. Hyb-Seq: Combining Target Enrichment and Genome Skimming for Plant Phylogenomics. *Applications in Plant Sciences* **2**: 1400042. doi: 10.3732/apps.1400042.

Wickham H. **2016**. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Table 1 – Genome size (2C), number of reads from GS (genome skimming) and raw TCS datasets, percentage of mapped reads on Bowtie2 Mapper and Geneious Read Mapper datasets, percentage of reads identified on "target clusters" of RepeatExplorer (RE) dataset and number of reads analysed for all datasets on each species.

Species	1C (Mbp)	Number of GS read pairs	Number of raw TCS reads [d]	% of filtered TCS reads*			Read pairs analysed
				BowTie2	Geneious	RE	
Rhynchospora cephalotes (L.) Vahl	356.97	11,737,291	6,932,932	11.77	23.12	16.80	600,000
Rhynchospora exaltata Kunth	244.5	21,384,261	5,151,874	13.02	20.65	18.45	422,199
Rhynchospora globosa (Kunth) Roem. & Schult.	-	4,000,000[c]	3,713,422	4.10	14.95	8.82	200,061
Rhynchospora pubera (Vahl) Boeckeler	1,613.7[a]	4,000,000[a]	4,164,982	3.96	12.29	8.66	2,797,919
Rhynchospora tenuis Link	381.42[b]	4,000,000[c]	4,826,084	10.18	20.64	6.36	661,128

[a]Marques *et al*., 2015; [b]Ribeiro *et al*,. 2018; [c]Ribeiro *et al*., 2017; [d]Buddenhagen 2016
*% of target reads idenfied by the different strategies

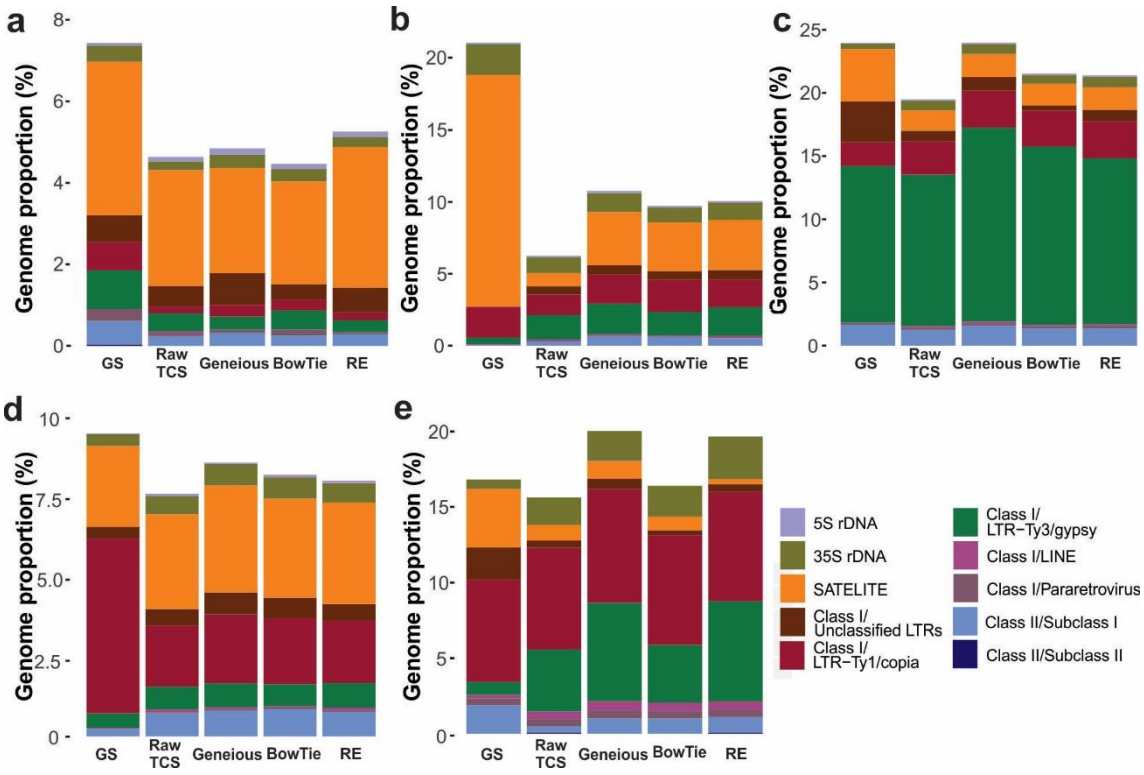


Figure 1 – Barplots representing genomic abundance of classified repeat types identified in every dataset (GS = Genome Skimming; Raw TCS = Raw Target Capture Sequencing; Geneious = Geneious filtered dataset; BowTie = BowTie2 filtered dataset; RE = RepeatExplorer custom database filtered dataset) of *Rhynchospora cephalotes* (a), *R. exaltata* (b), *R. globosa* (c), *R. pubera* (d) and *R. tenuis* (e). Bar colours represent different repeat types according to the caption at the lower right corner.

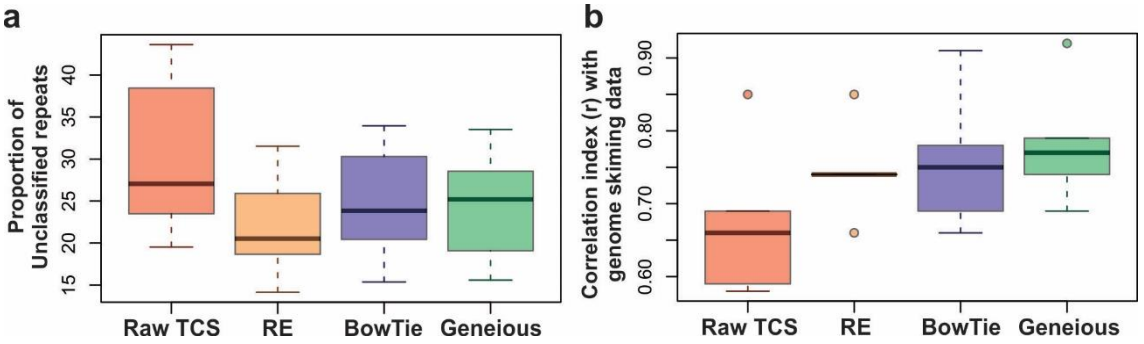


Figure 2 – Comparison of genomic proportions of unclassified elements (a) and correlation index (r) with genome skimming data (b) between raw and filtered Target Capture datasets (RepeatExplorer, BowTie and Geneious) of all *Rhynchospora* species.

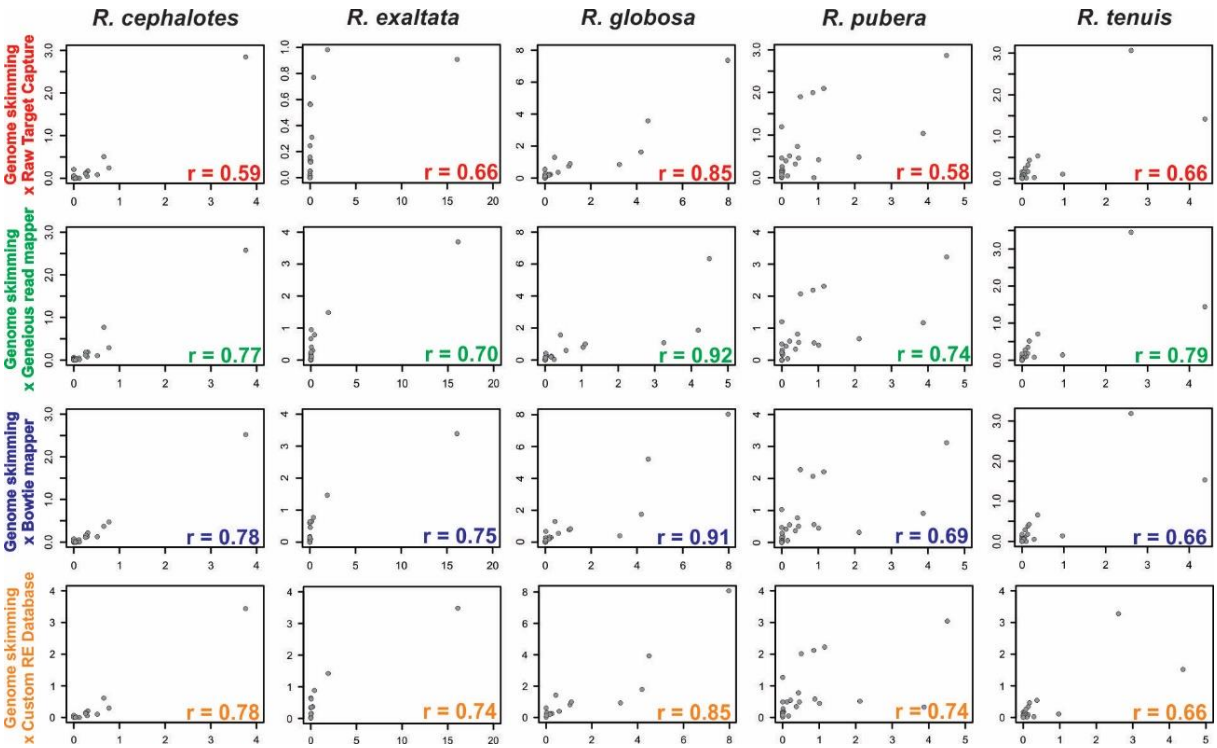


Figure 3 – Correlation between repeat abundances observed on genome skimming and target capture datasets of *Rhynchospora* species. Genome skimming abundance values are represented on the x axis of each plot, while target dataset abundances are on the y axis. Spearman's correlation index (r) for each case is plotted in the lower right corner.

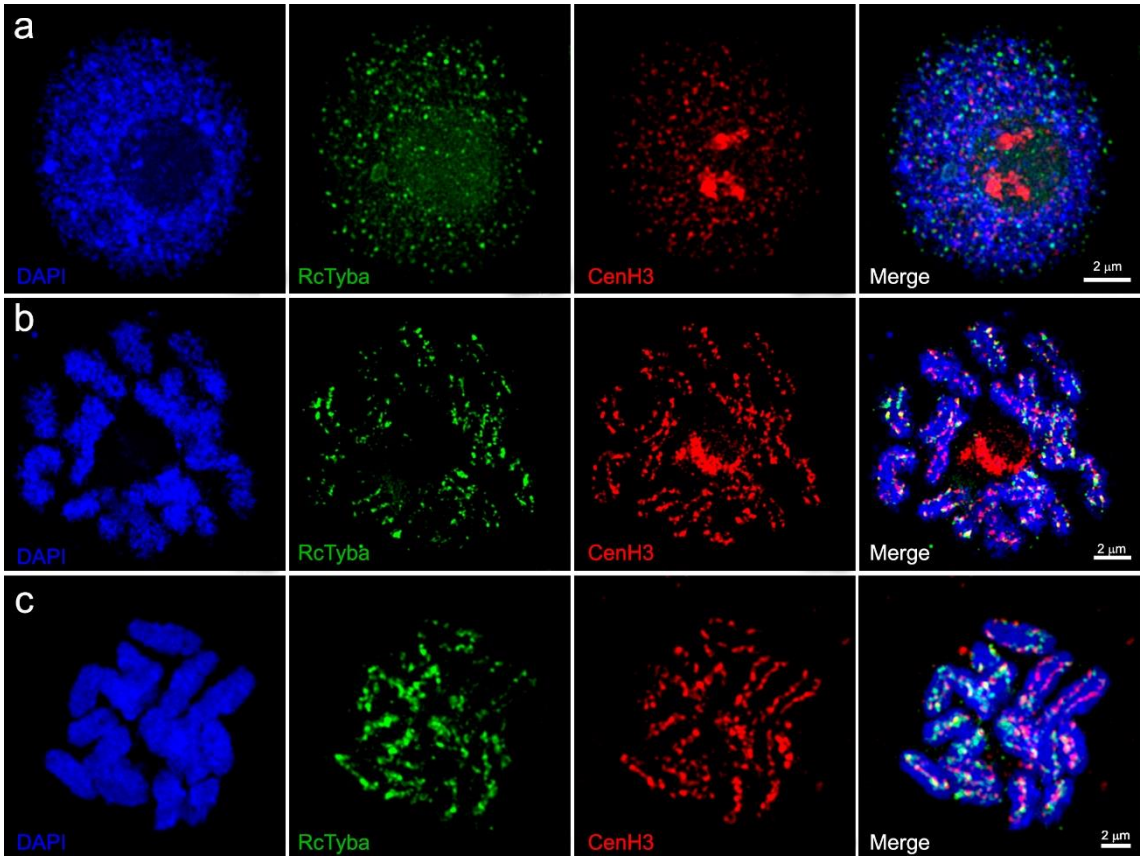


Figure 4 – Co-localization of *R. cephalotes* Tyba repeats and CENH3 in interphase nuclei (a), prometaphase (b) and metaphase chromosomes (c) via 3D-SIM imaging. Both Tyba and CENH3 clearly indicate the presence of holocentromeres in the condensed mitotic chromosomes.

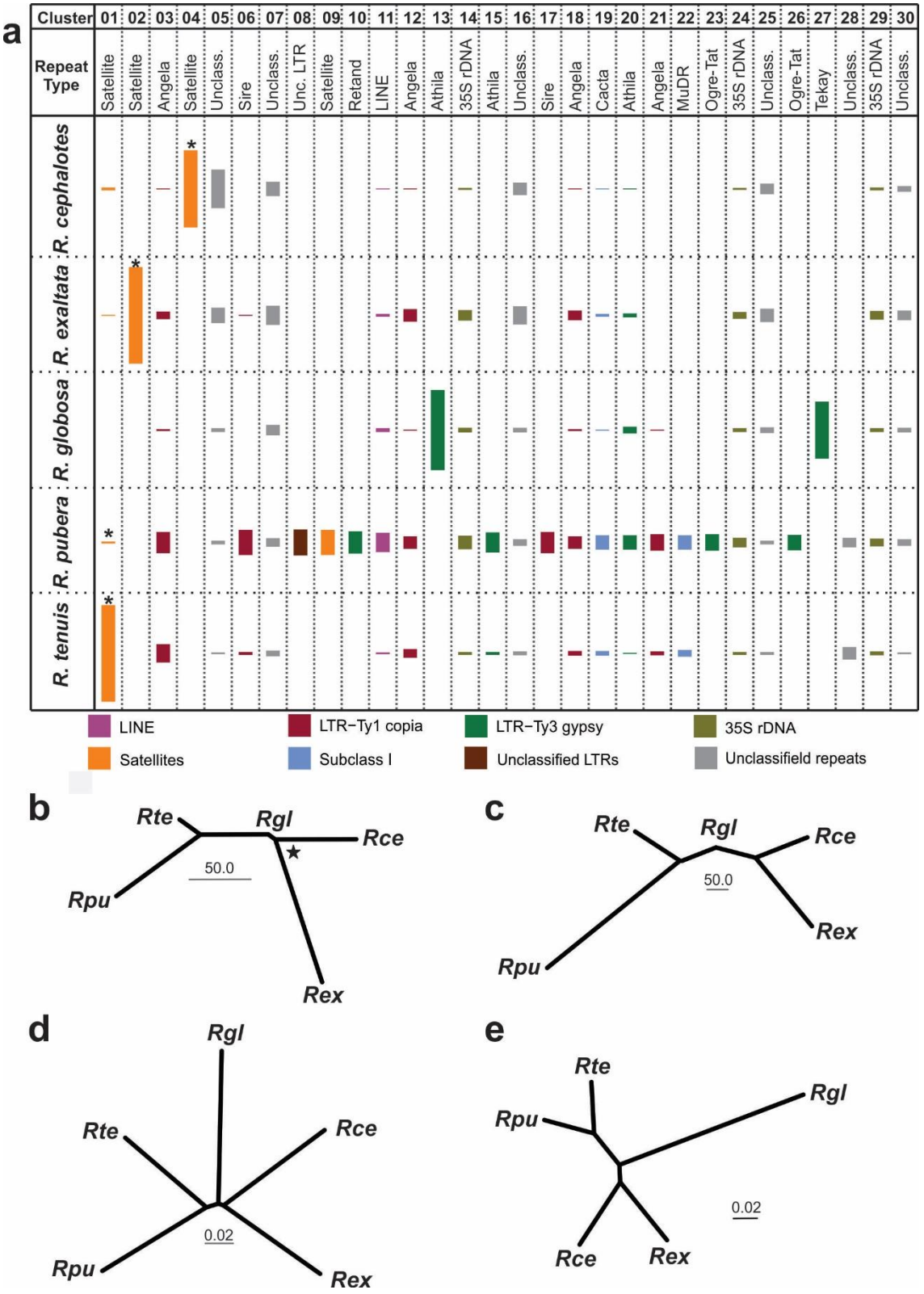


Figure 5 – Phylogenomics of Rhynchospora species using genome skimming and target capture datasets a) Graphic representation of the 30 most abundant clusters originated from the comparative clustering analysis with the BowTie dataset. Height of rectangles represent the genomic abundance of each cluster, with colours indicating the repeat type, according to the caption. Asterisks (*) indicate the clusters that represent Tyba, which is absent in Rhynchospora globosa. Below, phylogenetic trees obtained by repeat abundance of the GS (b) and BTM (c) datasets, AAF of total reads in repetitive clusters (d) and Bayesian inference based on 256 target regions (e). Star in b represents the only node with support below 100 (BS = 99).

Supplementary Table 1 Detailed annotation of repetitive elements at lineage level for all datasets on all *Rhynchospora* species.

R. cephalotes					
Repeat type/Lineage	GS	Raw TCS	Geneious	BowTie	RE
UNCLASSIFIELD	7.392404	38.44978	28.5435	30.28848	25.92024287
rDNA 5S	0.064996	0.10635	0.147251	0.12636	0.128520938
rDNA 35S	0.401899	0.210484	0.325503	0.301439	0.254776283
SATELLITE DNA	3.761397	2.844521	2.575951	2.520998	3.437111245
Unclassified LTR	0.656913	0.505674	0.769668	0.369095	0.611092346
Class I/LTR/Ty1/copia/ALE	0.306891	0.170262	0.186174	0.215879	0.205757079
Class I/LTR/Ty1/copia/ANGELA	0.115084	0	0.047534	0.054228	0
Class I/LTR/Ty1/copia/TORK	0.019876	0.0196	0.031172	0.018248	0.02368575
Class I/LTR/Ty1/copia/BIANCA	0.04174	0	0	0	0
Class I/LTR/Ty1/copia/IKEROS	0.515592	0.082319	0.103851	0.126188	0.099480149
Class I/LTR/Ty1/copia/IVANA	0	0	0.010506	0	0
Class I/LTR/Ty1/copia/SIRE	0	0.057095	0.066306	0.075231	0.068997619
Class I/LTR/Ty1/copia/TAR	0	0	0.033584	0	0
LTR-Ty3_gipsy/ATHILA	0.771202	0.243378	0.290025	0.470492	0.294115224
Class I/LTR/Ty3/gypsy/chromovirus/Reina	0	0	0.019117	0	0
Class I/LTR/Ty3/gypsy/non-chromovirus/RETAND	0	0.206224	0	0	0
Class I/LTR/Ty3/gypsy/chromovirus/TEKAY	0.052871	0	0	0	0
Class I/LTR/Ty3/gypsy/non-chromovirus/Tat/Ogre	0.121444	0	0.011539	0	0
Class I/non-LTR/Pararetrovirus	0.293375	0.045165	0.078534	0.127221	0.054580206
Class I/non-LTR/LINE	0	0.057095	0	0.014633	0
Class II/Subclass I/TIR/MuDR_MUTATOR	0.263759	0.124586	0.185141	0.157347	0.150558984
Class II/Subclass I/TIR/EnSpm_CACTA	0.258591	0.104987	0.112118	0.107251	0.126873234
Class II/Subclass I/TIR/hAT	0.059033	0.010396	0.027556	0	0.012563746
Class II/Subclass II/Helitron	0.034982	0	0	0	0
R. exaltata					
Repeat type/Lineage	GS	Raw TCS	Geneious	BowTie	RE
UNCLASSIFIELD	6.21209	43.62766	33.52828	33.95855	31.53302565
rDNA 5S	0.104813	0.099191	0.16343	0.110793	0.120246418
rDNA 35S	2.136043	1.10048	1.305278	1.042877	1.201011929
SATELLITE DNA	16.10136	0.907774	3.693127	3.390573	3.481337117
Unclassified LTR	0	0.5655	0.646252	0.596279	0.646251884
Class I/LTR/Ty1/copia/ALE	0.022321	0.049269	0.041825	0.018343	0.041824841
Class I/LTR/Ty1/copia/ALESIA	0	0.01417	0.015394	0.010028	0.015393865
Class I/LTR/Ty1/copia/ANGELA	1.91186	0.982322	1.421173	1.464772	1.421173245
Class I/LTR/Ty1/copia/IKEROS	0.200891	0.31	0.371486	0.650331	0.371485915
Class I/LTR/Ty1/copia/IVANA	0	0.051013	0.058961	0.096363	0.058961408
Class I/LTR/Ty1/copia/SIRE	0	0.028122	0.03224	0	0.032239982
LTR-Ty3_gipsy/ATHILA	0.036879	0.561358	0.618078	0.643727	0.618078207
Class I/LTR/Ty3/gypsy/chromovirus/CRM	0	0.01417	0.015394	0	0.015393865

Class I/LTR/Ty3/gypsy/chromovirus/Reina	0	0.245036	0.134478	0.04598	0.134478482
Class I/LTR/Ty3/gypsy/non-chromovirus/RETAND	0.399841	0.770205	0.885292	0.770662	0.885292469
Class I/LTR/Ty3/gypsy/chromovirus/TEKAY	0	0.133418	0.338665	0.170226	0.338665032
Class I/non-LTR/Pararetrovirus	0	0	0.014523	0	0.014522514
Class I/non-LTR/LINE	0.067934	0.129058	0.15452	0.09245	0.154519552
Class II/Subclass I/TIR/MuDR_MUTATOR	0.054347	0.118594	0.358706	0.46005	0.358706102
Class II/Subclass I/TIR/EnSpm_CACTA	0	0.158052	0.147549	0.147725	0.147548745

R. globosa

Repeat Type/Lineage	GS	Raw TCS	Geneious	BowTie	RE
UNCLASSIFIELD	14.69587	27.04666	25.19573	23.84443	20.51788651
rDNA 5S	0.037092	0.10617	0.102952	0.117128	0.116447201
rDNA 35S	0.447608	0.76382	0.792483	0.692716	0.837757591
SATELLITE DNA	4.189369	1.631293	1.862184	1.744358	1.789202971
Unclassified LTR	3.247036	0.844328	1.07623	0.391098	0.926058787
Class I/LTR/Ty1/copia/ALE	0.421543	1.296682	1.564878	1.294947	1.422201128
Class I/LTR/Ty1/copia/ALESIA	0	0.013083	0.012053	0	0.014348944
Class I/LTR/Ty1/copia/ANGELA	0.24661	0.220894	0.038168	0.293575	0.242276405
Class I/LTR/Ty1/copia/TORK	0	0.010567	0.038168	0.025638	0.011589532
Class I/LTR/Ty1/copia/IKEROS	1.041578	0.744196	0.799012	0.774656	0.816234175
Class I/LTR/Ty1/copia/IVANA	0.0406	0.166551	0.278223	0.187506	0.182673098
Class I/LTR/Ty1/copia/SIRE	0.174933	0.225926	0.232522	0.316699	0.24779523
LTR-Ty3_gipsy/ATHILA	7.973735	7.36296	8.451645	8.03913	8.0756962
Class I/LTR/Ty3/gypsy/chromovirus/CRM	0.01203	0.031197	0.040679	0.010054	0.034216713
Class I/LTR/Ty3/gypsy/chromovirus/GALADRIEL	0	0.547454	0	0	0.600448129
Class I/LTR/Ty3/gypsy/chromovirus/Reina	0	0.177621	0.026115	0.025135	0.194814512
Class I/LTR/Ty3/gypsy/non-chromovirus/RETAND	0.011027	0.110699	0.20239	0.269446	0.121414144
Class I/LTR/Ty3/gypsy/chromovirus/TEKAY	4.49312	3.576568	6.338357	5.211459	3.922780605
Class I/LTR/Ty3/gypsy/non-chromovirus/Tat/Ogre	0.021052	0.282281	0.417836	0.672609	0.309606066
Class I/LTR/Ty3/gypsy/non-chromovirus/Tat/TatII	0	0.027675	0.018582	0.017594	0.030353536
Class I/non-LTR/Pararetrovirus	0.162903	0.171583	0.195861	0.183987	0.188191923
Class I/non-LTR/LINE	0.01203	0.121768	0.128063	0.112604	0.133555558
Class II/Subclass I/TIR/MuDR_MUTATOR	1.098719	0.899677	1.003912	0.832969	0.986765858
Class II/Subclass I/TIR/EnSpm_CACTA	0.573419	0.360776	0.59612	0.547437	0.395699732

R. pubera

Repeat Type/Lineage	GS	Raw TCS	Geneious	BowTie	RE
UNCLASSIFIELD	9.865057	23.47654	19.08628	20.44017	18.66150942
rDNA 35S	0.607182	1.820938	1.988511	2.039564	2.818976485
SATELLITE DNA	3.869134	1.037677	1.167537	0.907774	0.326979111
Unclassified LTR	2.116154	0.483412	0.66771	0.314338	0.513091476
Class I/LTR/Ty1/copia/ALE	0.432153	0.732609	0.812793	0.768564	0.777588542
Class I/LTR/Ty1/copia/ALESIA	0.01456	0.121892	0.162537	0.132047	0.129375795
Class I/LTR/Ty1/copia/ANGELA	4.509153	2.865961	3.225208	3.117624	3.041921451
Class I/LTR/Ty1/copia/TORK	0.117306	0.394385	0.434813	0.417439	0.489493988

Class I/LTR/Ty1/copia/BIANCA	0.159643	0.043885	0.04527	0.05673	0.04657939
Class I/LTR/Ty1/copia/IKEROS	0.374532	0.320921	0.342963	0.366614	0.340624615
Class I/LTR/Ty1/copia/IVANA	0.025919	0.17148	0.193735	0.192649	0.182008454
Class I/LTR/Ty1/copia/SIRE	1.149722	2.092657	2.311076	2.206849	2.221139246
LTR-Ty3_gipsy/ATHILA	0.510735	1.897881	2.072398	2.275486	2.014404728
Class I/LTR/Ty3/gypsy/chromovirus/CRM	0.011772	0.259443	0.31231	0.289651	0.27537243
Class I/LTR/Ty3/gypsy/chromovirus/Reina	0	0.170224	0.233442	0.148795	0.180674683
Class I/LTR/Ty3/gypsy/non-chromovirus/RETAND	0	1.193401	1.199608	1.028688	1.266672139
Class I/LTR/Ty3/gypsy/chromovirus/TEKAY	0	0.460503	0.502991	0.467295	0.488775803
Class I/LTR/Ty3/gypsy/chromovirus/Tcn1	0	0.141708	0	0.056923	0.150408339
Class I/LTR/Ty3/gypsy/non-chromovirus/Phygy	0	0.114352	0	0	0.121373169
Class I/LTR/Ty3/gypsy/non-chromovirus/Tat/Ogre	0.852533	1.995318	2.187482	2.067154	2.117823286
Class I/LTR/Ty3/gypsy/non-chromovirus/Tat/TatII	0	0	0	0.045597	0
Class I/non-LTR/Pararetrovirus	0.45993	0.459923	0.552734	0.503889	0.488160217
Class I/non-LTR/LINE	0.215095	0.510477	0.593423	0.549098	0.541818853
Class II/Subclass I/TIR/MuDR_MUTATOR	0.882169	0	0.539098	0.557714	0.581421595
Class II/Subclass I/TIR/EnSpm_CACTA	1.009079	0.418937	0.467102	0.447353	0.44465876
Class II/Subclass I/TIR/hAT	0	0.015369	0	0	0.016313046
Class II/Subclass II/Helitron	0	0.067471	0	0	0.071613247

R. tenuis

Repeat Type/Lineage	GS	Raw TCS	Geneious	BowTie	RE
UNCLASSIFIELD	8.232814	19.52665	15.58929	15.35823	14.12850314
rDNA 5S	0.021192	0.077239	0.046696	0.078066	0.082544154
rDNA 35S	0.383001	0.57813	0.683254	0.692955	0.617835516
SATELLITE DNA	2.606601	3.06471	3.454221	3.186101	3.275192575
Unclassified LTR	0.373874	0.539899	0.70668	0.660298	0.539277398
Class I/LTR/Ty1/copia/ALE	0.133029	0.163493	0.180223	0.174016	0.083706747
Class I/LTR/Ty1/copia/ALESIA	0	0.052529	0.075431	0.073712	0.056136668
Class I/LTR/Ty1/copia/ANGELA	4.372952	1.419682	1.439753	1.530846	1.517184796
Class I/LTR/Ty1/copia/TORK	0	0.019582	0	0	0.020926687
Class I/LTR/Ty1/copia/BIANCA	0.100081	0.010413	0.097295	0	0.011127683
Class I/LTR/Ty1/copia/IKEROS	0.968948	0.101173	0.139774	0.135139	0.108121215
Class I/LTR/Ty1/copia/IVANA	0.060173	0.108943	0.165543	0.08973	0.116425456
Class I/LTR/Ty1/copia/SIRE		0.083611	0.120721	0.125497	0.089353631
LTR-Ty3_gipsy/ATHILA	0.139062	0.321546	0.346547	0.381312	0.343629485
Class I/LTR/Ty3/gypsy/chromovirus/Reina		0.068847	0.083084	0.067958	0.073575574
Class I/LTR/Ty3/gypsy/non-chromovirus/RETAND		0.156033	0.168822	0.143692	0.166749155
Class I/LTR/Ty3/gypsy/chromovirus/TEKAY	0.040992	0.152148	0.078086	0.064848	0.162597035
Class I/LTR/Ty3/gypsy/chromovirus/Tcn1	0	0	0	0	0.012907379
Class I/LTR/Ty3/gypsy/non-chromovirus/Tat/Ogre	0.290344	0.022224	0.081053	0.05334	0.023750129
Class I/LTR/Ty3/gypsy/non-chromovirus/Tat/TatII	0	0.014764	0.01062	0	0.015778058
Class I/non-LTR/Pararetrovirus	0	0.038231	0.03592	0.040588	0.040856865

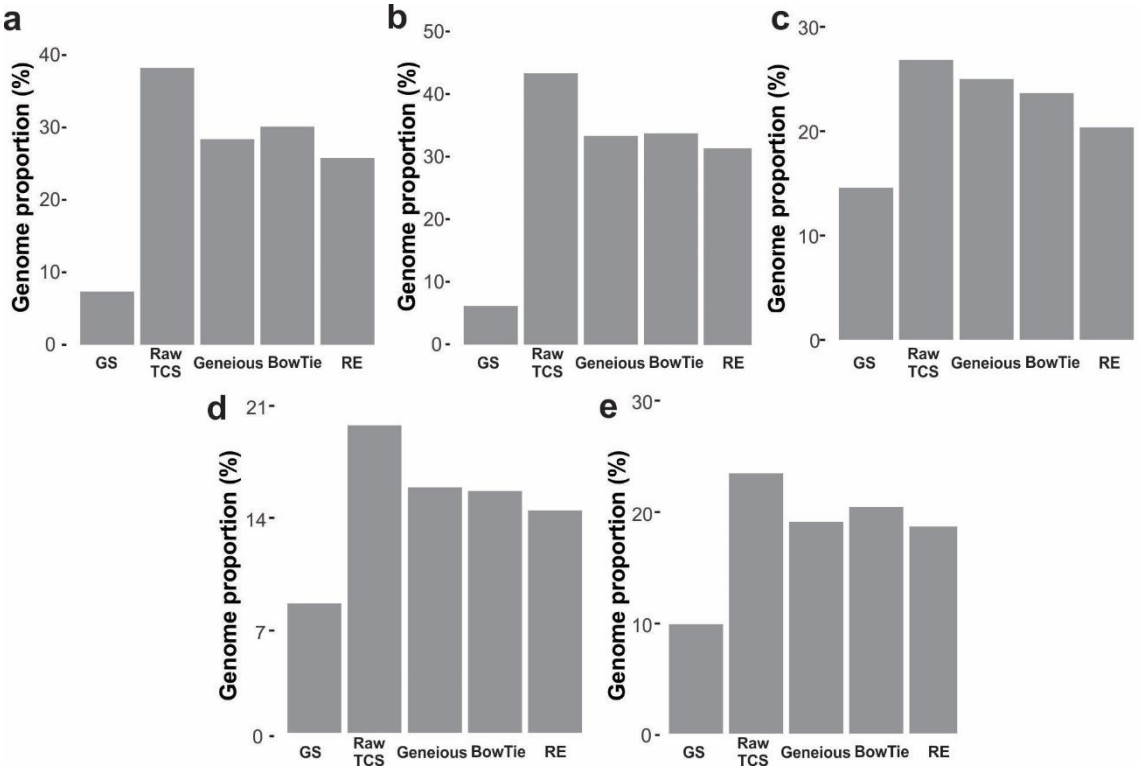
Class I/non-LTR/LINE	0.033412	0.081746	0.071839	0.05163	0.087360613
Class II/Subclass I/TIR/MuDR_MUTATOR	0.170773	0.435462	0.516931	0.422678	0.465369655
Class II/Subclass I/TIR/EnSpm_CACTA	0.072393	0.242597	0.278924	0.288472	0.259258398
Class II/Subclass I/TIR/hAT		0.046002	0.031391	0.173239	0.049161106
Class II/Subclass I/TIR/PIF_Harbinger	0	0.012277	0	0	0.0131207

Supplementary Table 2 Name and genomic abundance of the satellites found in all datasets of the five *Rhynchospora* species.

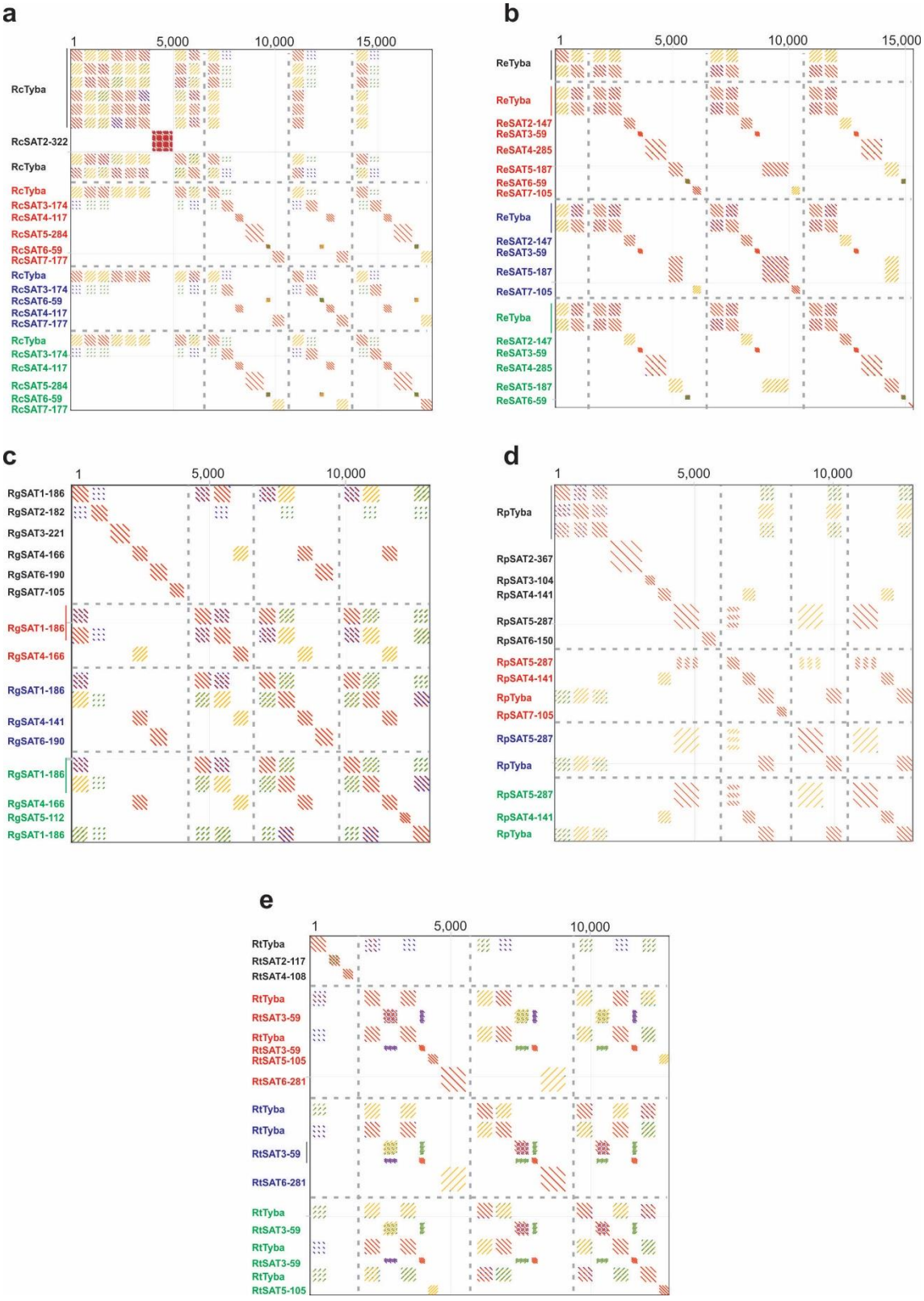
Sp.	Satellite	Genomic abundance (%)			
		GS	Raw TCS	BowTie	Geneious
R. cephalotes	*Rc*Tyba	3.66	2.04	2.38	2.37
	*Rc*SAT2-322	0.03	-	-	-
	*Rc*SAT3-174	-	0.07	0.07	0.08
	*Rc*SAT4-117	-	0.03	0.03	0.04
	*Rc*SAT5-284	-	0.03	-	0.04
	*Rc*SAT6-59	-	0.02	0.03	0.03
	*Rc*SAT7-177	-	0.01	0.01	0.01
R. exaltata	*Re*Tyba	16.24	2.31	2.73	2.92
	*Re*SAT2-147	-	0.24	0.37	0.29
	*Re*SAT3-59	-	0.19	0.20	0.24
	*Re*SAT4-285	-	0.14	-	0.19
	*Re*SAT5-187	-	0.03	0.03	0.04
	*Re*SAT6-59	-	0.01	-	0.01
	*Re*SAT7-105	-	0.01	0.01	-
R. globosa	*Rg*SAT1-186[a]	3.44	1.46	1.51	1.67
	*Rg*SAT2-182	0.24	-	-	-
	*Rg*SAT3-221	0.14	-	-	-
	*Rg*SAT4-166	0.12	0.07	0.07	0.08
	*Rg*SAT5-112	-	-	-	0.07
	*Rg*SAT6-190	0.05	-	0.05	-
	*Rg*SAT7-156	0.01	-	-	-
R. pubera	*Rp*Tyba[b]	3.21	0.12	0.14	0.14
	*Rp*SAT2-367	0.20	-	-	-
	*Rp*SAT3-104	0.11	-	-	-
	*Rp*SAT4-141	0.10	0.16	-	0.19
	*Rp*SAT5-287	0.09	0.67	0.70	0.75
	*Rp*SAT6-150	0.04	-	-	-
	*Rp*SAT7-105	-	0.01	-	-
R. tenuis	*Rt*Tyba[a]	2.56	2.66	2.77	2.97
	*Rt*SAT2-117	0.03	-	-	-
	*Rt*SAT3-59	-	0.38	0.40	0.44
	*Rt*SAT4-108	0.01	-	-	-
	*Rt*SAT5-105	-	0.01	-	0.01
	*Rt*SAT6-281	-	0.01	0.01	-

[a] – Previously described in Ribeiro et al. (2017)

[b] – Previously described in Marques et al. (2015)



Supplementary Figure 1 - Barplots of genomic abundance of unclassified repeats in every dataset of *R. cephalotes* (a), *R. exaltata* (b), *R. globosa* (c), *R. pubera* (d) and *R. tenuis* (e).



Supplementary Figure 2 - Dotplot comparison of all satellites found in the GS (genome skimming, black names), raw TCS (raw target capture, red names), BowTie (blue names), and Geneious (green names) of R. cephalotes (A), R. exaltata (B), R. globosa (C), R. pubera (D) and R. tenuis (E).