*Botanical Journal of the Linnean Society*, 2023, **XX**, 1–15
https://doi.org/10.1093/botlinnean/boad043
Advance access publication 25 August 2023
**Original Article**

# Original Article

# Genomic stability in *Cenostigma* Tul., (Caesalpinioideae, Fabaceae): causes and consequences

Natália Castro[1], Yennifer Mata-Sucre[1], Jefferson Carvalho-Sobrinho[2], André Marques[3], Rubens Teixeira de Queiroz[4], Gustavo Souza[1,*]

[1]Laboratório de Citogenética e Evolução Vegetal, Departamento de Botânica, Centro de Biociências, Universidade Federal de Pernambuco, Recife, PE, 50670-901, Brazil
[2]Federal University of Vale do São Francisco, Petrolina, PE, 56304-205, Brazil
[3]Max Planck Institute for Plant Breeding Research, Cologne 50829, Germany
[4]Department of Systematics and Ecology, Federal University of Paraíba, Exact and Nature Sciences Center, João Pessoa, PB, 58051-900, Brazil

[*]Corresponding author. Laboratório de Citogenética e Evolução Vegetal, Departamento de Botânica, Centro de Biociências, Universidade Federal de Pernambuco, Recife, PE, 50670-901, Brazil. E-mail: luiz.rodrigessouza@ufpe.br

## ABSTRACT

The Pantropical Caesalpinia group includes 225 species distributed in 27 monophyletic genera, among which *Cenostigma* stands out by taxonomic and phylogenetic complexity. The genus includes trees and shrubs with interspecific hybridization and high diversity in north-eastern Brazil (Caatinga domain). Detailed cytogenomic characterizations have been performed only in *C. microphyllum* revealing enrichment of long terminal repeats (LTR) Ty3/gypsy transposable elements (TEs) and satellite DNA (satDNA) in the heterochromatin. Here, we aimed to perform a comparative analysis of seven Northeast Brazilian species of *Cenostigma* using cytogenomic and genomic approaches. The comparative genomic analysis revealed repeats stability with similar TE abundance, composition, and chromosomal localization in all species. On the other hand, satDNA were highly variable in abundance, in some cases species-specific. Cytogenomic data confirmed the karyotype stability with the TE elements *Athila* and *Tekay* enriching the proximal heterochromatin. Moreover, the satDNA *CemiSat163* appeared to be exclusively located on acrocentric chromosomes of the analysed species. The genomic stability in *Cenostigma* may be related to their relatively recent age (~13.59 Mya), long-life cycle, and/ or similarity in ecological niche among this species. We propose that the genomic stability found in *Cenostigma* may facilitate the natural interspecific gene flow reported in sympatric species, complicating the interpretation of its systematics and evolution.

**Keywords:** Caatinga; Caesalpinia; cytogenomics; genomics; heterochromatin; repetitive elements; satellite DNA; transposable elements

## INTRODUCTION

The rapid evolution of repeats makes the repetitive fraction of the genome an important source of information for evolutionary studies of young plant lineages. High-throughput sequencing (HTS) technologies have recently emerged as a versatile source of genomics research for rapid access to different aspects of biodiversity (Dodsworth *et al.* 2019). Among the main HTS approaches, genome skimming is based on the sequencing (usually in low coverage) of small random genome fragments (reads) through Next-Generation Sequencing (NGS) technologies (e.g. Illumina, DNBseq). Among the bioinformatics tools used for repeat analysis using HTS, RepeatExplorer2 (https://repeatexplorer-elixir.cerit-sc.cz/galaxy) stands out as allowing a clustering approach to characterize the repetitive sequences in non-model genomes (Novák *et al.* 2013, 2020). This pipeline has been used to characterize repetitive genome fractions, discover new repetitive elements, and perform genomic comparative studies, thus contributing to the systematics of phylogenetically complex groups (Marques *et al.* 2015, 2018, McCann *et al.* 2020, Oliveira *et al.* 2021). When these repetitive elements have their chromosomal distribution determined it is possible to understand, for example, the composition of specific regions such as heterochromatin, as well as gain insight into their origin and evolution (González *et al.* 2018).

The 'Ecology of the Genome' concept was introduced to mirror the aspects of species populating an ecological

Received 29 November 2022; revised 30 May 2023; accepted 4 July 2023

community and repetitive lineages present in genomes (Brookfield 2005). By using this analogy, repeat dynamics can be studied from an ecological point of view, in which the genome can be compared to ecological communities, repeat lineages as species, and the copy numbers of a given repeat lineage as individuals (Schley *et al.* 2022). Therefore, it is possible to use ecological metrics, such as the Shannon diversity (Shannon 1948) and Simpson diversity index (Simpson 1949), to calculate the diversity of repeats (species) within a genome (ecological community). Although there are some papers using ecological methods to study genome dynamics (Brookfield 2005, Venner *et al.* 2009, Schley *et al.* 2022) few have directly addressed this topic, emphasizing the need to use these methods to conduct further research to better understand repeat dynamics.

Polymorphism in the repetitive elements seems to be associated with ecological conditions (Bilinski *et al.* 2018). Retroelements (RT) can undergo amplification/elimination influenced by ecological variables, especially stress conditions (e.g. temperature, precipitation, salinity) (Negi *et al.* 2016, Lyu *et al.* 2018). In this sense, the Caesalpinia group has stood out as an important model group to study genome–environment interaction. Heterochromatin distribution and composition, repetitive element abundance (Van-Lume *et al.* 2017, 2019, Mata-Sucre *et al.* 2020a), and genome size (Souza *et al.* 2019) were shown to be correlated with ecological variables, mainly temperature and latitude. The Caesalpinia group includes 205 species classified in 27 monophyletic genera, with a high diversity in the succulent biome (Gagnon *et al.* 2016, 2019). This group represents an ancient lineage (56 Mya) in which high niche conservatism and karyotypic stability of chromosome number $2n = 24$ (except a few polyploids) have been observed (Borges *et al.* 2012, Gagnon *et al.* 2019). On the other hand, an extensive heterochromatic variability in this group has been reported through the fluorochromes Chromomycin A3 (CMA) and 4′-6-diamidino-2-phenylindole (DAPI).

The CMA/DAPI banding patterns, described for 34 species in 10 Caesalpinia genera, revealed a correlation between heterochromatin patterns and the geographic distribution/ecological niche of the species (Van-Lume *et al.* 2017, Mata-Sucre *et al.* 2020a). Based on a genomic approach, Van-Lume *et al.* (2019) characterized repetitive fractions of Northeast Brazilian species of the Caesalpinia group [*Cenostigma microphyllum* (Mart. ex G.Don) Gagnon & G.P.Lewis, *Libidibia ferra* (Mart. ex Tul.) L.P.Queiroz, and *Paubrasilia echinata* (Lam.) Gagnon, H.C.Lima & G.P.Lewis]. The predominant long terminal repeats (LTR)-type TEs of the Ty3/gypsy superfamily (*Tekay* and *Athila*) were identified as the most abundant repeats in their genomes. In addition, species-specific satDNAs were characterized (Van-Lume *et al.* 2019). *In situ* hybridization revealed that all these repeats, except *Athila* in *C. microphyllum*, were enriched in the proximal heterochromatin CMA positive. However, intrageneric genomic diversity has not been estimated for any genus of the Caesalpinia group.

Within the Caesalpinia group, the Neotropical genus *Cenostigma* Tul., is composed of 14 species, with high diversity in Northeast Brazil, especially in the Caatinga domain where many of its species share a similar ecological niche and occur in sympatry (Fig. 1;Gagnon *et al.* 2016). The genus presents a relatively recent diversification within Caesalpinia group around 13.59 Myr (Gagnon *et al.* 2019). The phylogeny of *Cenostigma* is still poorly resolved and characterized by unsupported clades and low congruence within clades (Gagnon *et al.* 2016, 2019, Aecyo *et al.* 2021). In addition, recent studies using microsatellites and leaf morphometry demonstrated evidence for events of hybridization between *C. microphyllum* and *C. pyramidale* (Tul.) Gagnon & G.P.Lewis, in Caatinga areas (Aecyo *et al.* 2021). LTR-RTs are known to be activated in plant species following interspecific hybridization, this activation depends on the imbalance of TEs between the genomes of the parental species (Parisod *et al.* 2010, Usai *et al.* 2020). Thus, greater imbalance can lead to stronger genome shock effects during the hybridization process (Glombik *et al.* 2020). Therefore, characterizing the relative proportion of the repeat component between genotypes can be useful to better understand hybridization events.

Here we investigated the variation degree of repeats that compose the genomic repetitive fraction in *Cenostigma* species from north-eastern Brazil through a comparative cytogenomic approach. To characterize the global genomic repetitive fraction of seven *Cenostigma* species, we performed fluorescence *in situ* hybridization (FISH) using a set of specific repeat probes developed in *C. microphyllum* by Van-lume *et al.* (2019). Transfer of this probes was conducted for all seven *Cenostigma* species, together with additional chromosome number counting, CMA/DAPI double staining, 5S and 35S rDNA FISH. Considering the recent age of the genus and the similar ecological niche among species, we aimed to test the hypothesis that the heterochromatin composition and the presence/abundance of repeats are conserved among *Cenostigma* species.

## MATERIAL AND METHODS

### Plant material and genome sequencing

The species analysed here correspond to 50% of the 14 species from *Cenostigma* genus. All the species are Neotropical, with nine found in central and/or north-eastern Brazil, including parts of the Amazon (Gagnon *et al.* 2016). The species analysed in this study [*Cenostigma bracteosum* (Tul.) Gagnon & G.P.Lewis, *C. laxiflorum* (Tul.) Gagnon & G.P.Lewis, *C. macrophyllum* Tul., *C. microphyllum*, *C. nordestinum* Gagnon & G.P.Lewis, *C. pluviosum* (DC.) Gagnon & G.P.Lewis, and *C. pyramidale*] are distributed in Brazil, mostly in the Caatinga domain (Fig. 1). Five species of the genus occurs in the Mesoamerica: *C. eriostachys* (Benth.) Gagnon & G.P.Lewis, (Costa Rica and Panamá), *C. gaumeri* (Greenm.) Gagnon & G.P.Lewis, (Mexico and Cuba), *C. myabense* (Britton) Gagnon & G.P.Lewis, (Cuba), *C. pellucidum* (Vogel) Gagnon & G.P.Lewis, (Dominican republic), and *C. pinnatum* (Griseb.) E. Gagnon & G. P. Lewis (Cuba) (Fig. 1).

For the cytogenetic analysis, the seeds of each species used in this study were collected from natural and cultivated populations (Table 1). All seedlings were kept in the Experimental Garden of the Laboratory of Cytogenetics and Plant Evolution—UFPE. For repeatome analyses, we used previously published reads from *C. microphyllum* and *C. pyramidale* (GenBank accession number SRX11185454 and

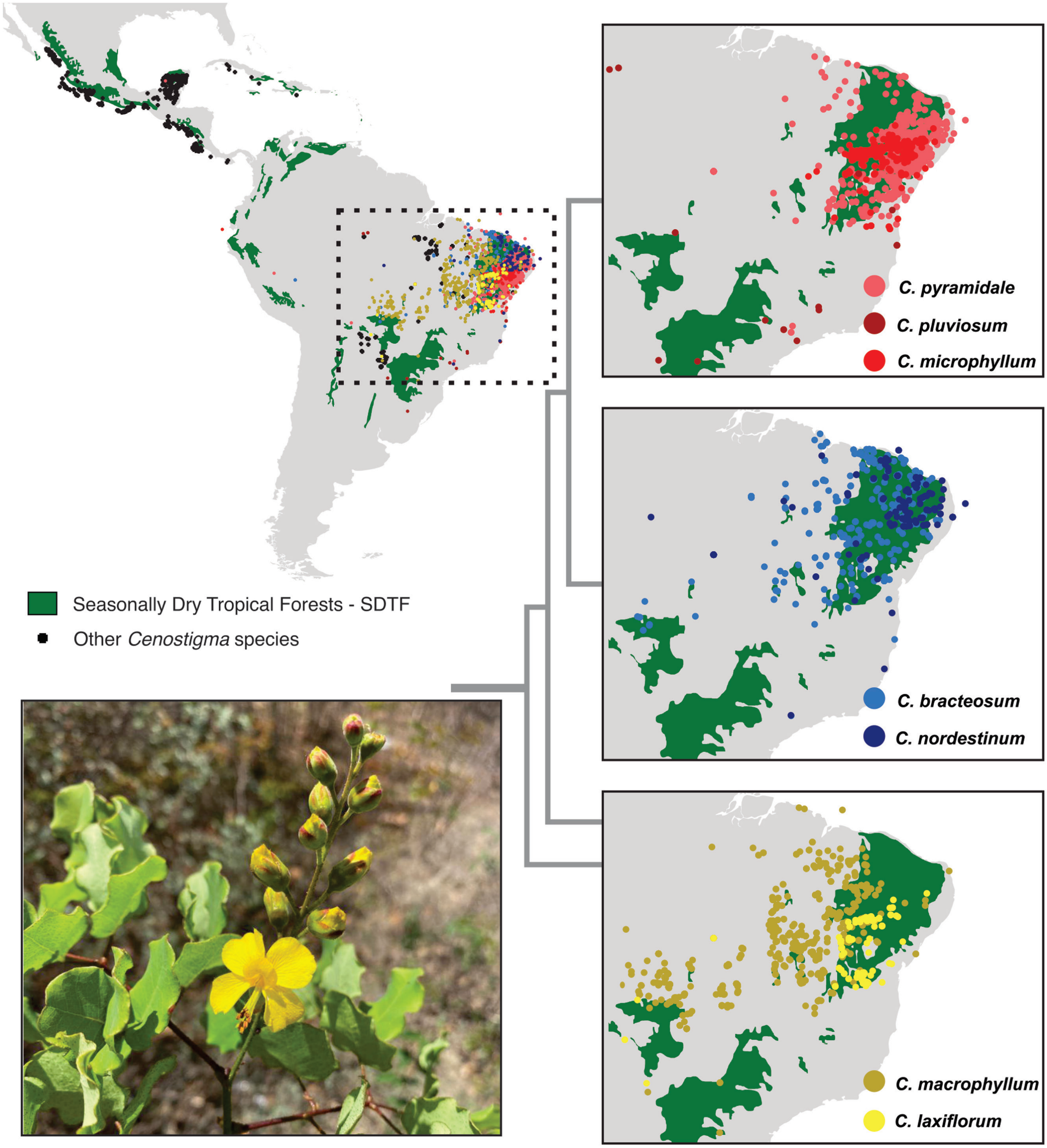

**Figure 1.** Distribution map of *Cenostigma* genus, focusing on the species from Northeast Brazil analysed here. Maps were organized based on their phylogenetic relationships: *C. pyramidale* + *C. pluviosum* + *C. microphyllum*, *C. bracteosum* + *C. nordestinum* and *C. macrophyllum* + *C. laxiflorum*. Featured, image of an inflorescence of *C. laxiflorum*.

SRX11185448, respectively) and performed new Illumina sequencing data (Illumina HiSeq 2000 platform) for *C. bracteosum*, *C. laxiflorum*, *C. macrophyllum*, *C. nordestinum*, and *C. pluviosum* by paired-end reads of 150 bp in a genome skimming approach (~0.1 × coverage; see genome sizes in Rodrigues *et al.* 2018, Souza *et al.* 2019). Information about the species analysed, voucher number, place of collection, and NCBI codes are available in Table 1.

## Flow cytometry

Absolute nuclear DNA contents were determined for *C. laxiflorum* and *C. nordestinum* by flow cytometry according to

Table 1. Voucher and NCBI codes of the *Cenostigma* ga species analysed in this study.

| Species | Provenance | Voucher | NCBI SRA code |
|---|---|---|---|
| *C. bracteosum* (Tul.) Gagnon & G.P.Lewis | Recife—PE, Brazil | Cultivated | SAMN34379857 |
| *C. laxiflorum* (Tul.) Gagnon & G.P.Lewis | Manoel Vitorino – BA, Brazil | UFP89958 | SAMN34379855 |
| *C. microphyllum* (Mart. ex G.Don) Gagnon & G.P.Lewis | Buíque – PE, Brazil | UFP88534 | SRX11185454 |
| *C. macrophyllum* Tul. | Petrolina—PE, Brazil | Cultivated | SAMN34379858 |
| *C. nordestinum* Gagnon & G.P.Lewis | Cabrobó – PE, Brazil | UFP89959 | SAMN34379859 |
| *C. pluviosum* (DC.) Gagnon & G.P.Lewis | Recife—PE, Brazil | Cultivated | SAMN34379856 |
| *C. pyramidale* (Tul.) Gagnon & G.P.Lewis | Buíque – PE, Brazil | UFP88533 | SRX11185448 |

Doležel *et al.* (2007). Fresh seeds (*C. laxiflorum*) or leaves (*C. nordestinum*) were collected to prepare the samples of 25–50 mg. The material was chopped together with fresh leaf tissue of the internal standard (*Raphanus sativus* 'Saxa', 1.11 pg/2C; Doležel *et al.* 1992) with a razor blade on a Petri dish (kept on ice) containing 1 mL of WPB isolation buffer (Loureiro *et al.* 2007). The solution was filtered through a 30 µm mesh filter and mixed with 50 µg/mL of propidium iodide (1 mg/mL).

Flow cytometry measurements were taken using a Partec Cyflow Space (Müster, Germany) equipped with a 488 nm laser canon. The relative fluorescence histograms were analysed on FloMax program v.2.3. The coefficient of variation of obtained peaks was assessed at half of the peak height (H.P.C.V.), discarding peaks with a H.P.C.V. > 5%. The genome size (ρg) of the samples were calculated using the following equation: 'sample DNA = (sample G1/standard G1) × standard DNA', where sample G1 is the peak position (G1) of the sample; standard G1 is the peak position (G1) of the standard, and standard DNA is the nuclear DNA (ρg) of the standard used in each measure. Three independent DNA estimations were performed for each sample. Measurements were exhausted with at least 1500 events per fluorescence peak.

### Chromosome spreads and fluorescent *in situ* hybridization (FISH)

Roots were pretreated with 0.002 M 8-hydroxyquinoline for 5 hours at 18°C. The material was then fixed in Carnoy (ethanol:acetic acid 3:1) and stored at −20°C until slide preparation. Chromosomal banding by double staining of the fluorochromes DAPI and CMA was performed according to Vaio *et al.* (2018). The FISH was used to locate repeat elements according to Pedrosa *et al.* (2002), with modifications. To localize the rDNA sites, 5S rDNA (D2) from *Lotus japonicus* (Regel) K. Larsen (Pedrosa *et al.* 2002) labelled with Cy3-dUTP (GE) and 35S rDNA (pTa71) (Gerlach and Bedbrook 1979) from *Triticum aestivum* labelled with Alexa-duTP (GE) were used as probes. Labelling of probes was done by nick translation. The repetitive DNA probes used in this study were designed and obtained by Van-Lume *et al.* (2019). The probes used were developed from the integrase domain of two TEs (Ty3/gypsy-*Tekay* and Ty3/gypsy-*Athila*) and from the most conserved region of the *Cemisat163* satDNA consensus sequences (detailed information of the probes can be found in Van-Lume *et al.* 2019). All probes were amplified from the genomic DNA of *C. microphyllum* and Sanger sequenced to confirm the protein domain, whereby complete families of the Ty3/gypsy-*Tekay* and Ty3/gypsy-*Athila* elements were mapped through FISH (see Van-Lume *et al.* 2019). Thus, *C. microphyllum* probes were used here as reference for the comparative cytogenomic analysis. Chromosomes were denatured at 75°C for 5 min. with the hybridization mixture containing formamide 50% (v/v), dextran sulphate 10% (w/v), 2 × SSC, and 50 ng/µL of each labelled probe. Stringent washes were performed, to give a final stringency of ~76%. The slides were hybridized with this mixture for at least 18 hours at 37°C. Finally, chromosomes were counterstained and mounted with DAPI/mountain medium. The best cells were analysed and captured with a Leica DMRB photomicroscope equipped with a Cohu CCD video camera using the Leica QFISH software. The images were edited using the Adobe Photoshop program.

### *In silico* analysis of the repetitive fraction

NGS data sequenced by the Illumina platform (2 × 150 bp) were used for the repetitive fraction analysis of the genome. The Galaxy/RepeatExplorer2 tool enabled a graph-based cluster analysis to identify the most abundant repetitive elements, grouping them based on similarity, and generating clusters for the different repetitive DNA families. The number of reads used as input for comparative analysis were adjusted to obtain 0.10 times coverage for each species [323 841 for *C. bracteosum* (1C = 449.9 Mbp); 692 715 for *C. laxiflorum* (1C = 1046 Mbp), 592 629 for *C. macrophyllum* (1C = 894.87 Mbp), 608 821 for *C. microphyllum* (1C = 919.3 Mbp), 579 470 for *C. nordestinum* (1C = 875.31 Mbp), 608 821 for *C. pluviosum* (1C = 919.3 Mbp), and 582 914 *C. pyramidale* (1C = 880.2 Mbp)] (Table 2). Only clusters with abundance above 0.01% were considered, this proportion was calculated from the number of clustered reads and the total number of reads used for the analysis. The chloroplast and mitochondria sequences were excluded, as they represent possible contaminants. The TAREAN (Tandem Repeat Analyser) pipeline, also implemented in Galaxy/RepeatExplorer2, was used to identify tandem repeats. This tool performs graph-based clustering analysis, allowing the identification and characterization of satDNA (Novák *et al.* 2017). The BLAST tool was used to characterize, when possible, the unidentified clusters by comparisons with custom and public databases (i.e. NCBI, https://www.ncbi.nlm.nih.gov/).

To analyse the homology of the most abundant repeats and better understand the dynamic evolution of these

**Table 2.** Genomic composition (expressed as a percentage) of repetitive sequences in *C. bracteosum*, *C. laxiflorum*, *C. macrophyllum*, *C. microphyllum*, *C. nordestinum*, *C. pluviosum*, and *C. pyramidale* by cluster analysis. Genome size are expressed as Mbp.

| **Repetitive DNA** | **Order** | **Superfamily** | **Lineages/Class** | ***C. bracteosum* (1C = 449.9 Mbp)** | ***C. laxiflorum* (1C = 1046.46 Mbp)** | ***C. macrophyllum* (1C = 894.87)** | ***C. microphyllum* (1C = 919.3 Mbp)** | ***C. nordestinum* (1C = 875.31 Mbp)** | ***C. pluviosum* (1C = 919.3 Mbp)** | ***C. pyramidale* (1C = 880.2 Mbp)** |
|---|---|---|---|---|---|---|---|---|---|---|
| Total reads | | | | 323 841 | 692 715 | 592 629 | 608 821 | 579 470 | 608 821 | 582 914 |
| Coverage | | | | 0.10× | 0.10× | 0.10× | 0.10× | 0.10× | 0.10× | 0.10× |
| Retro transposons | LTR | | | | | | | | | |
| | | Ty3/Gypsy | Chromovirus/ *Tekay* | 14.79 | 10.14 | 7.87 | 9.15 | 12.11 | 10.12 | 9.33 |
| | | | Chromovirus/ CRM | 0.25 | 0.43 | 0.54 | 0.35 | 0.52 | 0.42 | 0.40 |
| | | | Non-chromovirus/ OTA/ Tat/ Ogre | 1.37 | 1.36 | 1.51 | 0.83 | 1.28 | 0.81 | 1.09 |
| | | | Non Chromovirus/ OTA/*Athila* | 2.79 | 3.35 | 2.83 | 2.53 | 3.40 | 2.60 | 3.07 |
| | | Ty1/Copia | SIRE | 0.65 | 0.85 | 0.10 | 0.44 | 1.06 | 0.44 | 0.67 |
| | | | Ale | 0.44 | 0.43 | 0.49 | 0.30 | 0.51 | 0.38 | 0.44 |
| | | | Tork | 0.28 | 0.50 | 0.46 | 0.22 | 0.42 | 0.27 | 0.33 |
| | | | Bianca | 0.18 | 0.36 | 0.19 | 0.21 | 0.32 | 0.28 | 0.30 |
| | | | Ikeros | 0.14 | 0.23 | 0.09 | 0.15 | 0.19 | 0.18 | 0.17 |
| | | | Ivana | 0.10 | 0.18 | 0.14 | 0.13 | 0.19 | 0.16 | 0.19 |
| | | | TAR | 0.21 | 0.20 | 0.21 | 0.13 | 0.17 | 0.18 | 0.16 |
| | Non-LTR | | LINE | 0.51 | 0.39 | 0.29 | 0.33 | 0.49 | 0.28 | 0.33 |
| DNA transposon | | | Mutator | 0.23 | 0.28 | 0.64 | 0.26 | 0.25 | 0.27 | 0.33 |
| | | | Pararetrovirus | 0.15 | 0.27 | 0.13 | 0.17 | 0.20 | 0.19 | 0.20 |
| | | | Harbinger | 0.01 | 0.02 | 0.04 | 0.02 | 0.02 | 0.03 | 0.01 |
| | | | hAT | 0.00 | 0.01 | 0.00 | 0.01 | 0.02 | 0.01 | 0.01 |
| | | | CACTA | 0.03 | 0.06 | 0.08 | 0.04 | 0.04 | 0.04 | 0.04 |
| rDNA | | | 35S/18S | 7.62 | 3.76 | 5.65 | 2.83 | 3.55 | 2.41 | 2.85 |
| | | | 5S | 0.78 | 0.23 | 1.03 | 0.10 | 0.31 | 0.09 | 0.20 |
| Total satDNA | | | | 4.65 | 1.43 | 1.50 | 2.43 | 2.13 | 0.98 | 1.70 |
| | | | CemiSat163 | 3.69 | 0.86 | 1.35 | 1.92 | 1.60 | 0.58 | 0.97 |
| | | | CepySat222 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.12 |
| Unclassified | | | | 5.83 | 6.18 | 8.47 | 4.80 | 5.86 | 5.43 | 6.08 |
| Shannon index | | | | 1.91 | 2.29 | 2.18 | 2.23 | 2.18 | 2.12 | 2.20 |
| Total | | | | 42.11 | 32.10 | 33.75 | 26.15 | 34.44 | 26.53 | 28.94 |

elements, all sequences belonging to the Ty3/gypsy-*Tekay* and Ty3/gypsy-*Athila* lineages were used for comparative approaches. The reverse transcriptase protein domains of Ty3/gypsy-*Tekay* and Ty3/gypsy-*Athila* lineages from *C. bracteosum*, *C. laxiflorum*, *C. macrophyllum*, *C. microphyllum*, *C. nordestinum*, *C. pluviosum*, and *C. pyramidale* were extracted and filtered for quality (alignment sequence identity 0.35, alignment similarity 0.45, and alignment ratio length 0.8) from contigs using DANTE (domain-based annotation of transposable elements) on the RepeatExplorer2 platform (Novák *et al.* 2020). This tool annotates and classifies protein domains on the basis of homology comparisons with the available Viridiplantae protein domain database (Neumann *et al.* 2019). All reverse transcriptase protein obtained for Ty3/gypsy-*Tekay* and Ty3/gypsy-*Athila* from different clusters in RE (considered here as different lineages of the same element) were aligned together using MAFFT (Katoh and Standley 2013). For the satDNA *CemiSat163*, the monomers representing the consensus sequences of the satDNA present in the contigs of *C. bracteosum*, *C. laxiflorum*, *C. macrophyllum*, *C. microphyllum*, *C. nordestinum*, *C. pluviosum*, and *C. pyramidale* were aligned together using MAFFT. The alignment of each element for all species was used to construct Neighbor Joining phylogenetic trees using FastTree in Geneious Prime (v.7.1.9) (http://www.geneious.com) (Kearse *et al.* 2012).

To compare the diversity of repeats in the analysed species, Shannon's index (Shannon 1948) was used. This index is commonly used to measure the diversity of a given population or community. To calculate the Shannon index of the genomes, we first identified the repeat lineages in our study based on the lowest hierarchical classification in the REXdb plant repeat database (Neumann *et al.* 2019). These lineages were treated as 'species' within a 'community' (genome). We then used the diversity() function of the Vegan (Oksanen *et al.* 2013) package from R software (R Core Team 2019) to calculate the Shannon index, as outlined by Schley *et al.* (2022). The abundance of all lineages was taken into consideration, providing us with a diversity measure for each genome.

### Comparative idiogram constructions and phylogenetic relationships

The comparative cytogenomic idiograms were drawn using Corel X7 and used for comparative interpretations following the phylogenetic relationships. Metaphases of each species showing clear chromosome morphology were measured using the Drawid (v.0.26) Program (Kirov *et al.* 2017). The largest metacentric and acrocentric pairs were used to represent the TEs and satDNA FISH distribution in each species. Evolutionary relations were performed using complete plastomes. The plastome of *C. microphyllum* available at the NCBI (MZ441392; Aecyo *et al.* 2021) was used as a reference to assemble the plastomes of *C. bracteosum*, *C. laxiflorum*, *C. macrophyllum*, *C. nordestinum*, *C. pluviosum*, and *C. pyramidale* (Table 1), more detailed information about the plastome macroevolution within the Caesalpinia group are available at Aecyo *et al.* (2021). The raw Illumina reads were mapped against the reference plastome (MZ441392) using Geneious (v.9.1.8). The alignments were made using MAFFT. For simplification, we used the most general model of DNA substitution GTR + I + G (Abadi *et al.* 2019) on full plastome alignment. Phylogenetic relationships were inferred using maximum likelihood (ML) with 1000 replicates in Geneious (v.9.1.8) using the FastTree (Price *et al.* 2009) plugin.

### Geographical distribution of *Cenostigma* species

To gather occurrence information of the 14 *Cenostigma* species, occurrence data 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 Coordinate Cleaner (Zizka *et al.* 2019) implemented in the R software (R Core Team 2019). Coordinate Cleaner allowed us to remove duplicate records. The occurrence maps were constructed using QGIS software (v.3.4.2) (http://qgis.osgeo.org), the shapefile used to plot the seasonally dry tropical forest regions was downloaded from the DRYFLOR website (http://www.dryflor.info/data) (Dryflor *et al.* 2016). Finally, all the maps were edited and formatted using Corel X7.

## RESULTS

### Comparative genome analysis

RepeatExplorer comparative *in silico* analysis included a total of 1 143 965 reads from seven species, i.e. *C. bracteosum*, *C. laxiflorum*, *C. macrophyllum*, *C. microphyllum*, *C. nordestinum*, *C. pluviosum*, and *C. pyramidale* (Table 2; Fig. 2). DNA content were estimated for the first time to *C. laxiflorum* (1C = 1.07 ± 0.05 pg; 1046.46 Mbp) and *C. nordestinum* (1C = 0.89 ± 0.03 pg; 875.31 Mbp). The repetitive fraction proportion in all genomes was relatively similar, being 27.70% (*C. microphyllum*), 28.65% (*C. pluviosum*), 31.05% (*C. pyramidale*), 33.20% (*C. laxiflorum*), 33.95% (*C. nordestinum*), 35.79% (*C. macrophyllum*), and 37.93% (*C. bracteosum*) (Supporting Information, Table S1). The retrotransposons *Tekay*, *Athila* and *Ogre* were the most abundant elements (Table 2). On the other hand, satDNA showed a differential abundance among them (0.98% to 4.65%). Additionally, the *CemiSat163* showed a higher proportion in *C. bracteosum* (3.69%), while in the other species ranged from 0.58% to 1.92%. Notably, the satDNA cluster 193, named *CepySat222*, was found only in *C. pyramidale* genome (0.15%). The family diversity of these repeats was measured using Shannon's index, which was very similar for all genomes analysed, ranging from 1.91 to 2.29 (Table 2). Abundance values of each element in the individual analyses are presented in Supporting Information, Table S1.

The phylogenetic trees using the different contigs of the LTR-RT elements *Tekay*, *Athila* and the satDNA *CemiSat163* revealed a low rate of divergence between transcriptase domains in all the species (Supporting Information, Fig. S1). Moreover, alignments between the reverse transcriptase of the *Athila* and *Tekay* repeats lineages in all species revealed high similarity between them (Supporting Information, Fig. S2), with a sequence identity equal to 75.1% (Supporting Information, Fig. S2A) and

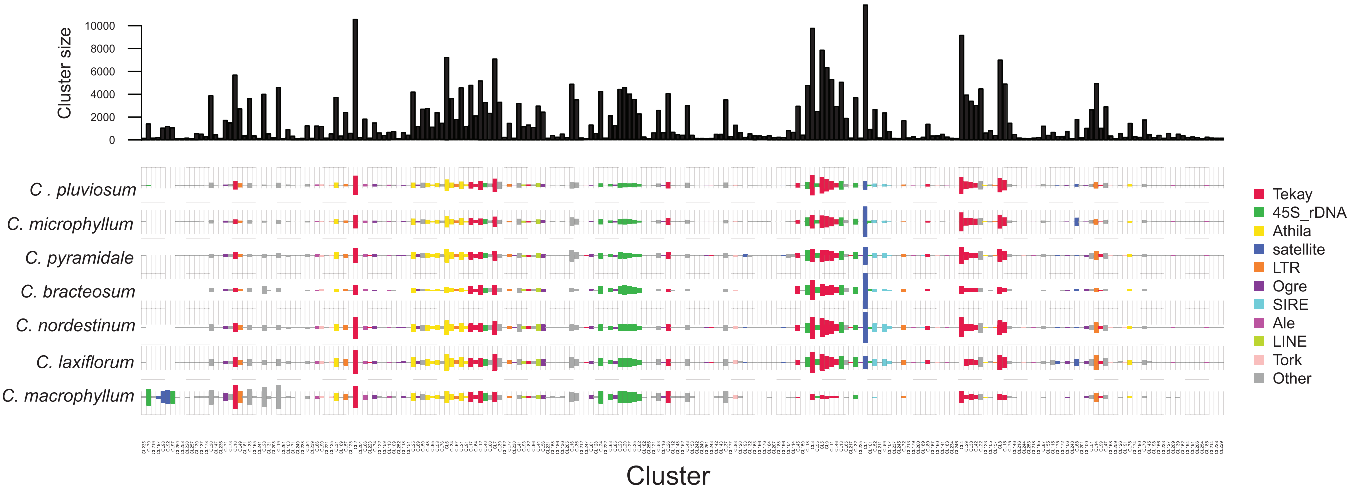

**Figure 2.** Comparative analysis by clustering showing the difference between all clusters of the repetitive elements present in the genomes of *Cenostigma pluviosum, C. microphyllum, C. pyramidale, C. bracteosum, C. nordestinum, C. laxiflorum*, and *C. macrophyllum*. The bar graph at the top shows the size of the individual clusters. The size of the rectangles in the bottom panel is proportional to the number of reads in each cluster for each species. The proportions of each cluster were adjusted using the genome size of each specie. The clusters and species were sorted using hierarchical clustering. Each colour in the rectangle indicates a different repeat lineage.

79.0% (Fig. S2B), respectively. Likewise, alignments between the satDNA revealed a high similarity, with a sequence identity equal to 71.4% (Supporting Information, Fig S3).

### Distribution of heterochromatin and 5S-35s rDNA sites in *Cenostigma* species

All the seven species analysed showed a numerical stable karyotype (2*n* = 24), with 16 metacentric/submetacentric chromosomes and eight acrocentric chromosomes (Figs 3–5). CMA$^+$/DAPI-positive bands were visualized in the proximal regions of all chromosome in all samples. For *C. laxiflorum, C. macrophyllum*, and *C. nordestinum,* we present the first characterization of the heterochromatin pattern (Fig. 3).

rDNA sites were analysed on *C. laxiflorum, C. macrophyllum*, and *C. nordestinum* chromosomes. 5S rDNA signals were observed in the terminal region of one acrocentric chromosome pair, which is adjacent to one 35S rDNA sites. Additional 35S rDNA sites were observed in three other pairs of acrocentric chromosomes (Fig. 3). The 35S rDNA fully occupied the short arms of the three acrocentric chromosome pairs (Fig. 3).

### Comparative cytogenomic analysis and phylogenetic relationships

Probes of the most abundant elements in the *C. microphyllum* genome (Van-Lume *et al.* 2019) were used for the comparative cytogenomic analyses of *C. bracteosum, C. laxiflorum, C. macrophyllum, C. nordestinum, C. pluviosum*, and *C. pyramidale* (Figs 4, 6). The TE *Tekay* was hybridized in *C. bracteosum, C. laxiflorum, C. macrophyllum, C. nordestinum, C. pluviosum*, and *C. pyramidale* and showed signals in all chromosomes of these six species with an enrichment in pericentromeric heterochromatin, colocalized with the CMA$^+$/DAPI-positive bands (Figs 4, 6, Supporting Information Fig. S4). These signals varied in intensity and in some cases with small signals that were difficult to visualize (Fig. 4, Supporting Information Fig. S4).

The TE *Athila* was hybridized in *C. laxiflorum, C. macrophyllum, C. pluviosum*, and *C. pyramidale*, with the addition of *C. microphyllum* (Figs 4, 6, Supporting Information Fig. S5) revealing signals in the proximal heterochromatin in almost all chromosomes of these five species, always colocalized with CMA$^+$/DAPI$^-$ bands (Figs 4, 6). Signals varied in size and intensity among the species analysed, as reported for the *Tekay* element. The chromosomal distribution of the *Athila* element in *C. microphyllum* revealed clear signals in the proximal heterochromatin of the entire chromosomal complement of this species, which differed from what we have previously reported as showing signals in only a few chromosome pairs (Van-Lume *et al.* 2019).

Signals obtained from *CemiSat163* hybridizations were unique to the acrocentric chromosome pairs and in variable numbers (Figs 5, 6). For the species *C. pluviosum, C. laxiflorum* and *C. macrophyllum*, bands in the terminal regions of the four acrocentric chromosome pairs were observed, with small variations in intensity among the species analysed. However, *C. laxiflorum* showed *CemiSat163* signals in the long arms of the first and second acrocentric pairs that were not colocalized with the CMA$^+$/DAPI$^-$ bands (Fig. 5). Additionally, *C. pyramidale* revealed signals in three of the four acrocentric chromosome pairs and the absence of the terminal signal in the long arm of the first acrocentric pair (Fig. 5). Phylogenetic analysis revealed that *Cenostigma* species analysed here form a high-support monophyletic group (posterior probability = 1; Fig. 6), corroborating previous analyses (Gagnon *et al.* 2016, 2019). The species are subdivided into three main subclades and although some species are phylogenetically distant, all showed karyotypic similarity in terms of heterochromatic bands, repeat distribution, and abundance (Fig. 6).

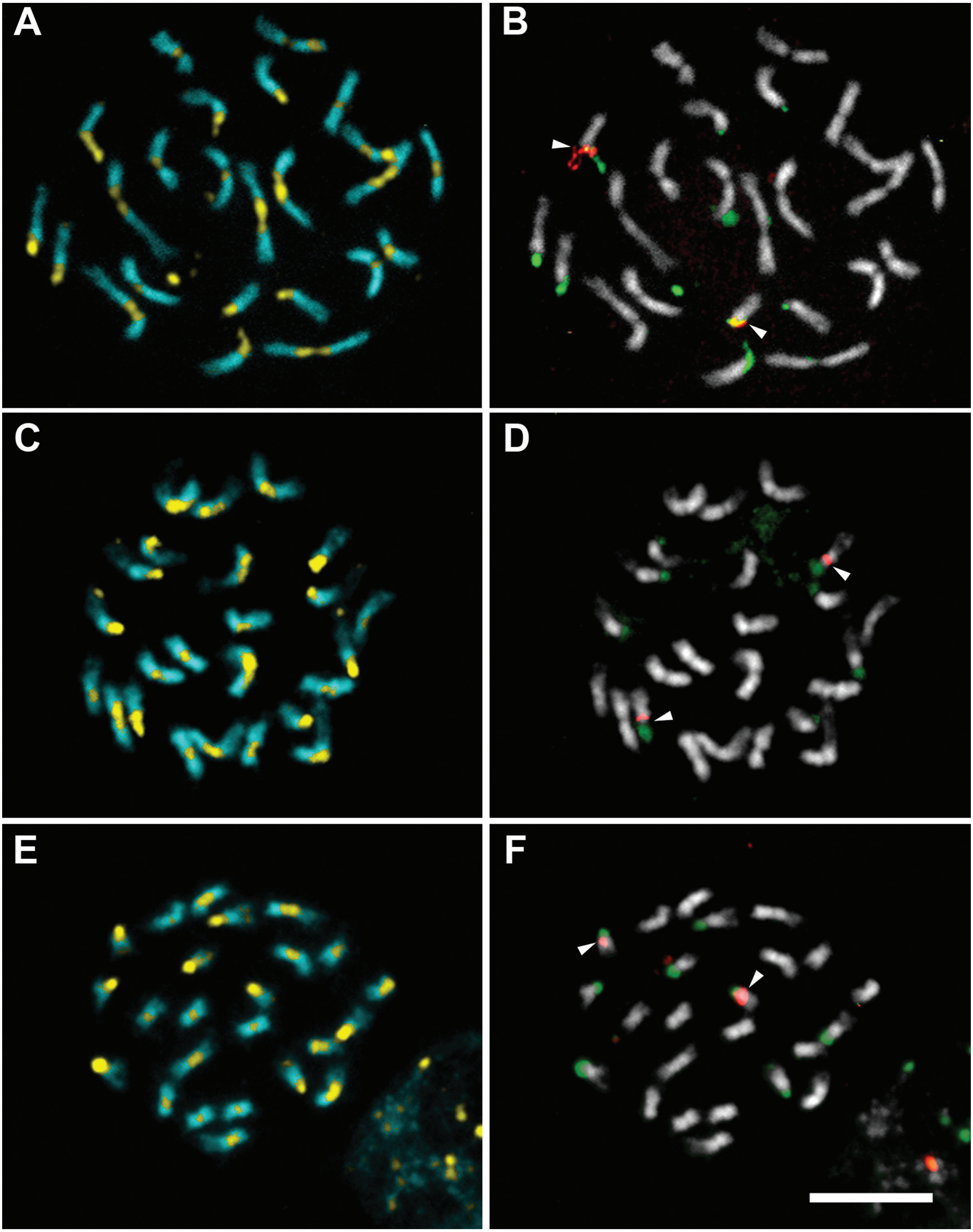

**Figure 3.** Double-stained DAPI (blue) and CMA (yellow) heterochromatic banding pattern and fluorescence *in situ* hybridization mapping of 5S (red) and 35S (green) rDNA in A, B, *C. laxiflorum*; C, D, *C. macrophyllum* and E, F, *C. nordestinum*. Arrows indicate the 5S adjacent to the 35S rDNA. The chromosomes are counterstained with DAPI (pseudocoloured in grey). Scale bar: 10 µm.

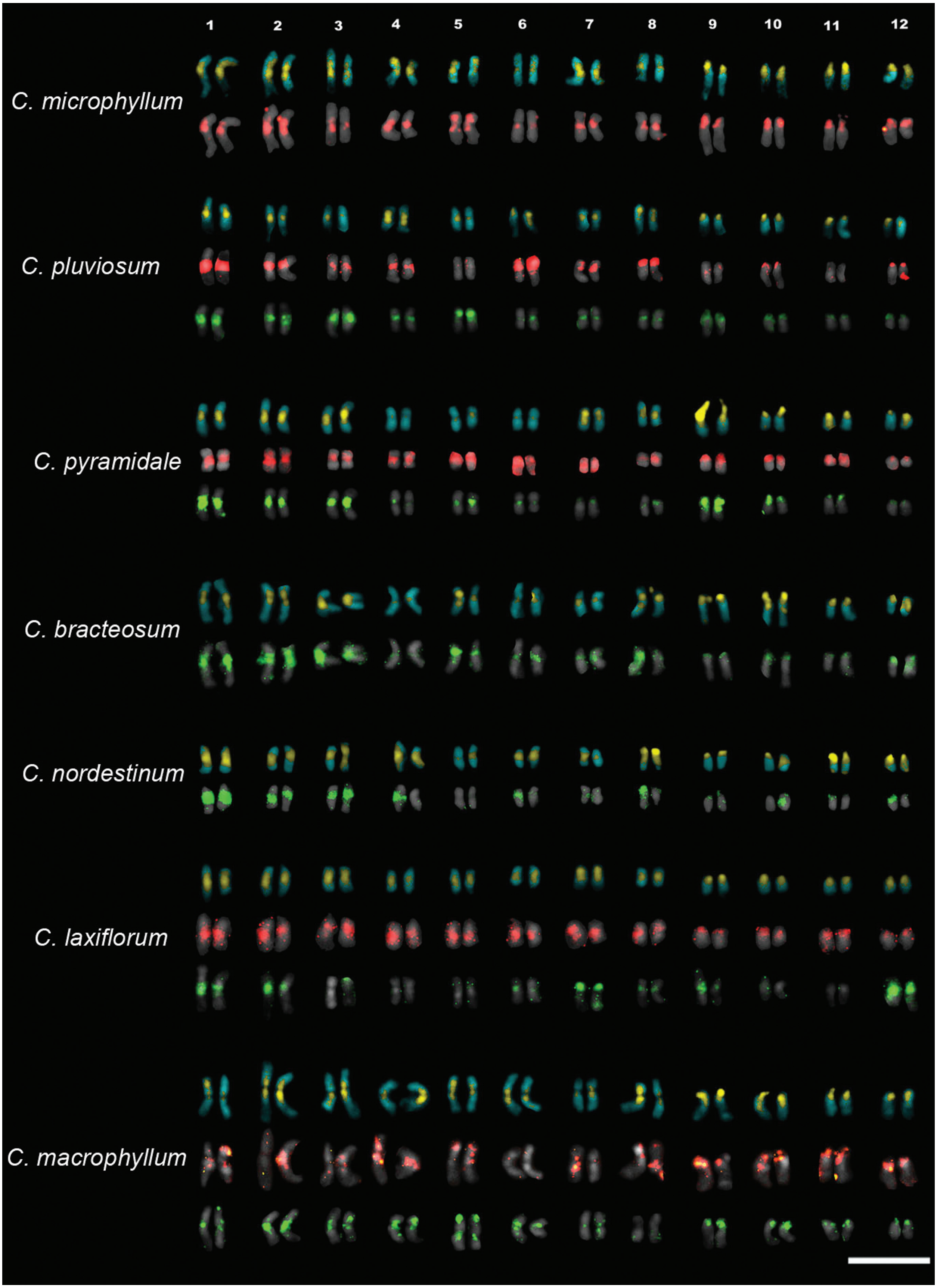

**Figure 4.** Karyogram of the 12 chromosome pairs of *Cenostigma microphyllum*, *C. pluviosum*, *C. pyramidale*, *C. bracteosum*, *C. nordestinum*, *C. laxiflorum*, and *C. macrophyllum* showing LTR elements signals. CMA/DAPI bands are shown in yellow/blue. The species-specific probes for TEs Ty3/gypsy-*Tekay* and Ty3/gypsy-*Athila* are labelled with Cy3-dUTP and pseudocoloured in green and red, respectively. The chromosomes are counterstained with DAPI (pseudocoloured in grey). Scale bar: 10 µm.

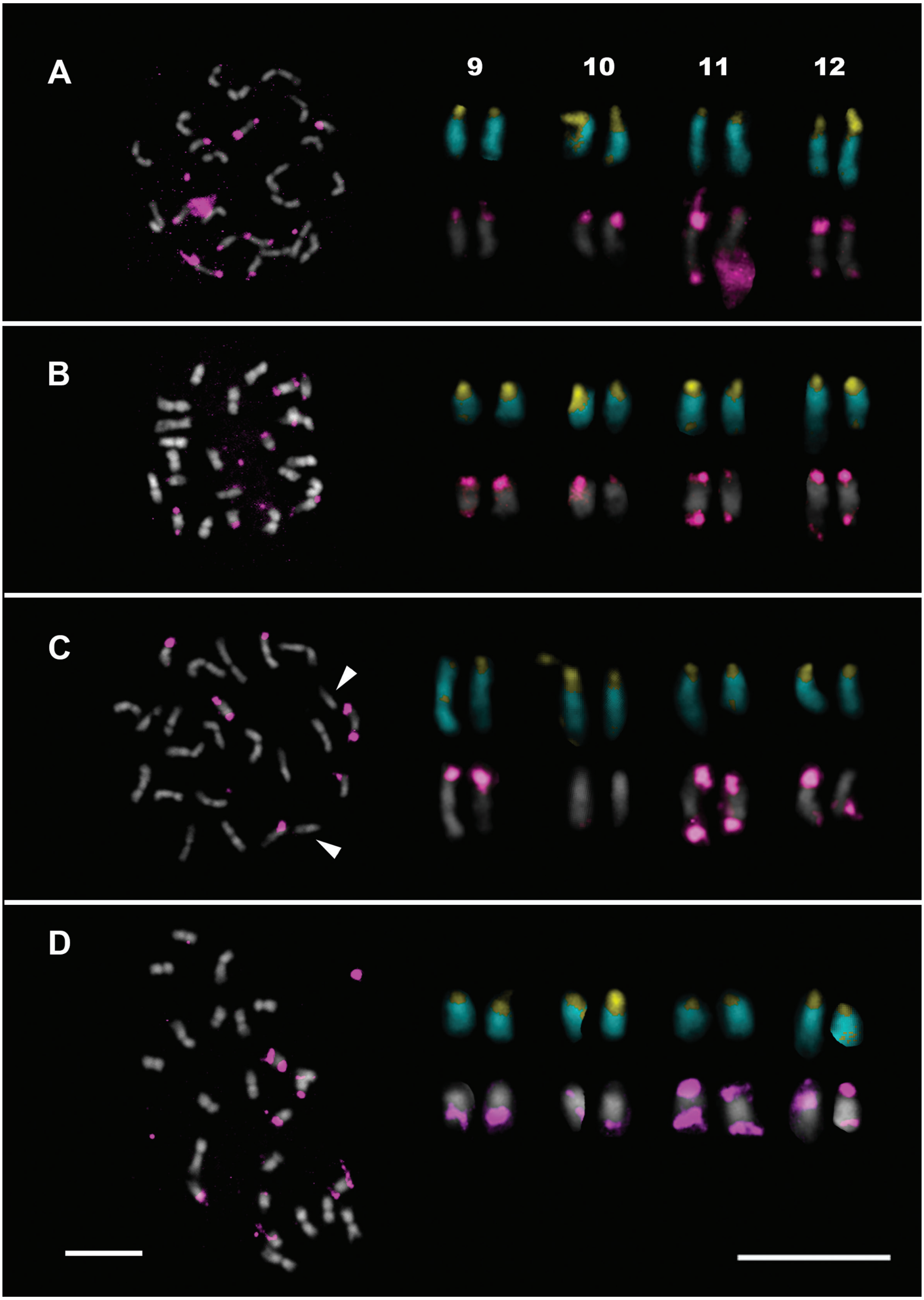

**Figure 5.** Scheme of acrocentric chromosomes of A, *Cenostigma pluviosum*; B, *C. macrophyllum*; C, *C. pyramidale*; and D, *C. laxiflorum* showing *CemiSat163* labelled with Cy3-dUTP pseudocoloured in purple and CMA$^+$/DAPI$^-$ positive bands in yellow. The chromosomes are counterstained with DAPI (pseudocoloured in grey). Arrows indicate the chromosome pairs that are lacking the satellite DNA signal. Scale bar: 10 µm.

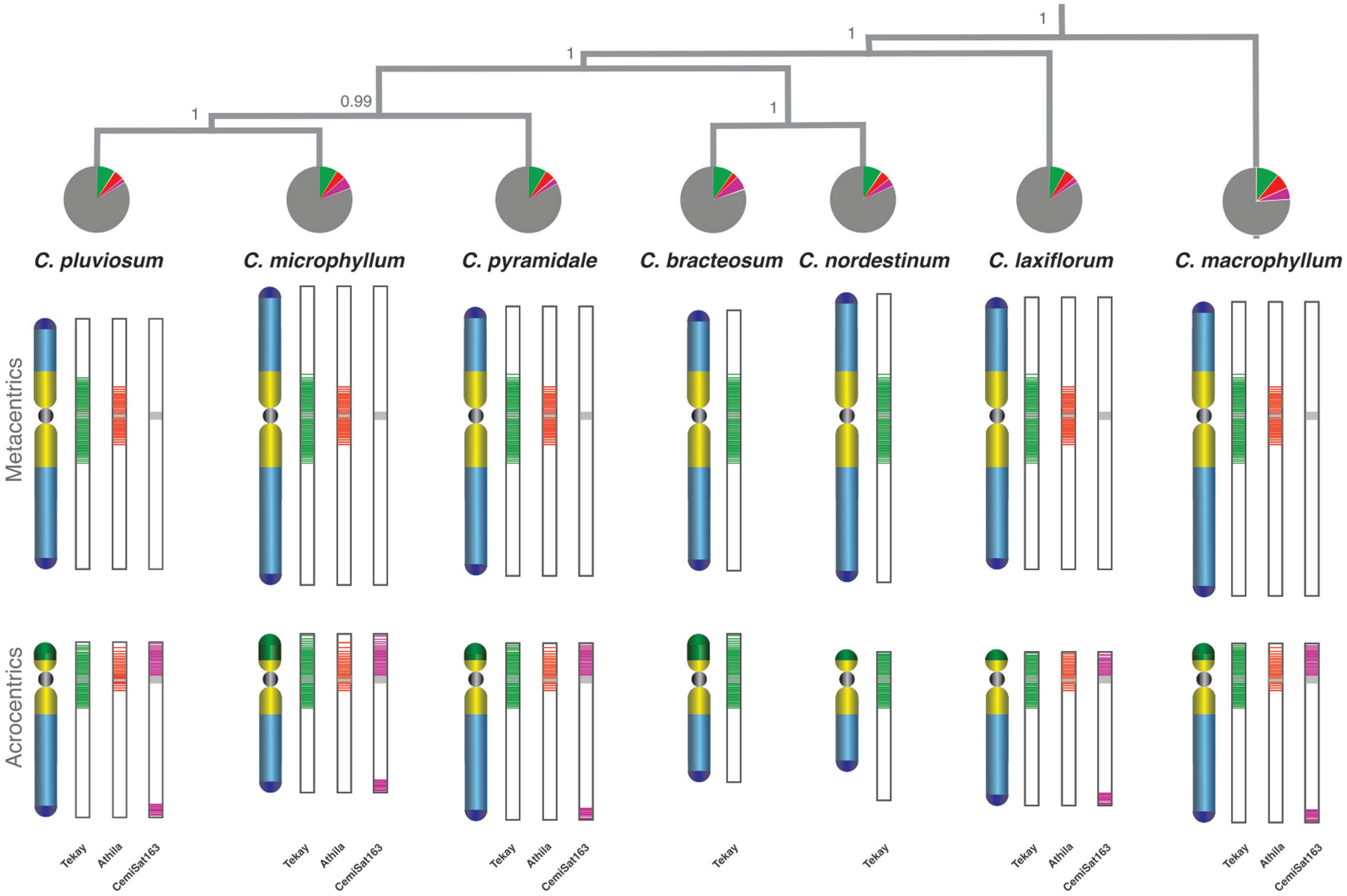

**Figure 6.** Comparative cytogenomic idiograms showing the cytogenetic distribution of repetitive sequences retrotransposons (TE *Tekay* and *Athila*), satDNA (*CemiSat163*) and 35S rDNA (green) in relation to constitutive heterochromatin in *Cenostigma pluviosum*, *C. microphyllum*, *C. pyramidale*, *C. bracteosum*, *C. nordestinum*, *C. laxiflorum*, *and C. macrophyllum*. *C. microphyllum* Tekay hybridization results were based on Van-Lume *et al.* (2019). CMA/DAPI banding are represented in yellow and blue. The phylogenetic topology is based on the maximum likelihood plastome tree. Distribution of repeats in the comparative cytogenomic idiograms were based on fluorescence intensity.

## DISCUSSION

### The karyotypic stability in the genus *Cenostigma*

The remarkable karyotypic stability (chromosome number, heterochromatic banding pattern, 5S/35S rDNA sites) reported here corroborates previous cytogenetic analyses of the genus *Cenostigma* (Van-Lume *et al.* 2017, 2019). A constant chromosome number of 2*n* = 24 was confirmed for seven of the 14 species of the genus, with an only description of polyploidy in *C. bracteosum* (2*n* = 4*x* = 48) that is the most significant cytogenetic difference reported for this genus (Alves and Custodio 1989). At the intergeneric level, the Caesalpinia group stands out for the variety of heterochromatin patterns (Van-Lume *et al.* 2017, Mata-Sucre *et al.* 2020a). However, the heterochromatic stability at generic level has been reported in other Caesalpinia group genera, such as *Arquita* Gagnon, G.P.Lewis & C.E.Hughes (proximal bands CMA0/DAPI-), *Coulteria* Kunth (proximal bands CMA0/DAPI-), *Libidibia* Schltdl., (proximal bands CMA$^+$/DAPI$^-$), and *Tara* Molina (proximal bands CMA0/DAPI-) (Van-Lume *et al.* 2017, Mata-Sucre *et al.* 2020a). By contrast, the genus *Erythrostemon* shows high karyotypic diversity, in terms of heterochromatin and genome sizes (Van-Lume *et al.* 2017, Souza *et al.* 2019, Mata-Sucre *et al.* 2020b). The karyotypic conservation in *Cenostigma* may be related to environmental conditions, since in Caesalpinia group, taxa that occur in similar ecological niches tend to show more conserved karyotypes (Gagnon *et al.* 2019, Van-Lume *et al.* 2019, Mata-Sucre *et al.* 2020a). To date, all *Cenostigma* species with available cyto-molecular analyses are from the Brazilian Northeast (Van-Lume *et al.* 2017, 2019, Mata-Sucre *et al.* 2020a).

The rDNA probes are typically used to characterize karyotypes, especially in cytotaxonomy (Dias *et al.* 2020, Nguyen *et al.* 2021). In angiosperms, rDNA show a non-random arrangement related to factors such as the amount of active mobile elements in the karyotype (Raskina *et al.* 2004). Particularly, the 35S rDNA is preferentially distributed in the terminal regions, especially in the short arms of acrocentric chromosomes (Raskina *et al.* 2008, Roa and Guerra 2012). Unlike most other genera in the Caesalpinia group, *Cenostigma* exhibits a conserved rDNA site distribution with synteny of the 5S and 35S site in one chromosomal pair (Van-Lume *et al.* 2017). This synteny of rDNA sites is homoplasic in the Caesalpinia group occurring in phylogenetically unrelated genera, such as *Coulteria*, *Mezoneuron* Desf. and *Tara* (Van-Lume *et al.* 2017, Mata-Sucre *et al.* 2020a).

### Similarity in heterochromatin composition in the genus *Cenostigma*

Our data showed that heterochromatin in *Cenostigma* is highly conserved and enriched by the *Tekay* and *Athila* TEs (Figs 4, 6). In addition, the comparative genomic analysis, showed in general a strong presence of LTR-TEs in the genomes of *C. bracteosum*, *C. laxiflorum*, *C. macrophyllum*, *C. microphyllum*, *C. nordestinum*, *C. pluviosum*, and *C. pyramidale*. LTR elements are particularly abundant in plants, especially the Ty3/gypsy superfamily, which is usually enriched in plant heterochromatic regions (Neumann *et al.* 2019).

The *Tekay* TE, which belongs to the Chromovirus subfamily, stands out in the genome of other Caesalpinia species as one of the most abundant and enriched element in the proximal heterochromatin (Van-Lume *et al.* 2019), as observed in the genus *Cenostigma.* Chromoviruses are characterized by its association with centromeric regions and presence of chromodomains known as CHDCRs (Neumann *et al.* 2019). The *Tekay* element, specifically, contains the type II chromodomain (i.e. CHDII), which plays a role in directing the *Tekay* insertion predominantly into peri/centromeric heterochromatin, as it is able to recognize and bind to methylated histone residues H3K9 (Jacobs and Khorasanizadeh 2002, Nielsen *et al.* 2002) and H3K27 (Fischle *et al.* 2003, Min *et al.* 2003) that are common in heterochromatic regions (Gao *et al.* 2008, Neumann *et al.* 2019). The TE *Athila*, was previously described in *C. microphyllum* showing terminal bands distribution on acrocentric chromosomes, colocalized with the satDNA *CemiSat163* (Van-Lume *et al.* 2019). However, our results revealed that this element presents an enrichment also in the proximal heterochromatin of all chromosomes, and not only in acrocentric chromosomes. This difference may be due to probe sensitivity and/or signal detection after *in situ* hybridization (Van-Lume *et al.* 2019). Additionally, a proximal distribution of the *Athila* element has also been reported for *Libidibia ferrea* corroborating its presence in this chromosome region (Van-Lume *et al.* 2017, 2019). Both elements, *Tekay* and *Athila* exhibit a predominantly clustered signals (forming chromosome bands) on *Cenostigma* chromosomes, which supports the typical TE distribution observed in other Caesalpinia genera (Van-Lume *et al.* 2019, Mata-Sucre *et al.* 2020a). This clustered pattern of TEs chromosomal distribution was also observed in other angiosperms families such as Cyperaceae (de Souza *et al.* 2018) and Poaceae (Topalian *et al.* 2022). Despite the rapid evolution of repetitive DNAs, even in phylogenetically close groups (Macas *et al.* 2015), we demonstrate here that the TEs that make up the heterochromatin of *Cenostigma* species are highly conserved in terms of chromosomal location, repeat diversity, and the sequences of the protein domains.

The most abundant satDNA in *C. microphyllum*, *CemiSat163* showed signals colocalized with CMA$^+$/DAPI$^-$ bands in the acrocentric chromosomes and small dot-like signals in the terminal region of these same chromosomes, without association with CMA$^+$ bands. The number, intensity, and size of these *CemiSat163* bands vary among *Cenostigma* species, with a constant number of at least three chromosomes with the same pattern and similar abundance in the genome ~1%. In this sense, despite the conservation of *CemiSat163* abundance in all analysed species, the absence of signal in an acrocentric pair of *C. pyramidale* may be a differential cytological mark for this species. The abundance of satDNA sequences using RE has been reported to be underestimated in some groups probably due to the use of short reads and/or inaccuracy of this method for quantifying tandem repeats (Ribeiro *et al.* 2020, Costa *et al.* 2021). In addition, signal intensity during *in situ* hybridization depends on detection sensitivity and/or spatial resolution, so the presence of minor signals is not discarded (De Jong *et al.* 1999). Thus, further analysis of the *Cenostigma* satellitome may help to characterize species-specific sequences that allow differentiation of the karyotypes of the genus.

The rapid evolution of the repetitive DNA makes this genomic fraction an important source for understand plant systematics and evolution. Differences in the rate of TE evolution may be related to environmental condition (Schley *et al.* 2022), plant habit (He *et al.* 2020) and the length of life cycles between perennial and annual species (Mascagni *et al.* 2017). As expected, the diversification of satDNA was greater than that of TEs, suggesting that this class of repetitive elements have differential rates of evolution (Lower *et al.* 2018). The proliferation of TEs and changes in its abundance may be related to recombination events whereby homologous chromosomes with differential numbers of elements co-segregate in subsequent generations (Mascagni *et al.* 2017). The occurrence of both processes can be related to the number of generations which the accumulation or loss can occur, which is higher in annual plants than in perennials. Therefore, annual genera (e.g. *Phaseolus* L.,) can present high diversification rates of satDNA and mobile elements (Ribeiro *et al.* 2020), even if they are recent (~5 Myr old, Delgado-Salinas *et al.* 2006). However, perennial genera with a longer life cycle, as *Cenostigma* (~13.59 Myr old, Gagnon *et al.* 2019), low heterogeneity of repeats can be found, as observed here. This trend related to plant habit, added to the high conservatism of the ecological niche of *Cenostigma* in Caatinga, may explain the stability in the repetitive fractions of the genomes observed here.

### Does genomic stability facilitate hybridization in *Cenostigma*?

Within the genus *Cenostigma*, the available morphological, phylogenetic and genomic evidences suggest the existence of natural hybrids (Lewis 1995, Aecyo *et al.* 2021). Our phylogenetic analyses are better resolved than the literature (Gagnon *et al.* 2019) however, more robust phylogenetic analyses should be performed to better understand *Cenostigma* evolution. Lewis (1995) proposed the existence of interspecific hybrids in *Cenostigma* based on the difficulty in separating species by morphological characters. In addition, phylogenetic analyses reveal low resolution and non-monophyly of some species (Gagnon *et al.* 2016, Aecyo *et al.* 2021). Hybridization in plants, are more common in younger lineages as well as potential introgression with other species (Mallet 2005), so it seems to be a correlation between the recent age of a lineage and permeability of reproductive barriers. There is evidence that the expansion and diversification of the genus *Cenostigma* in the Neotropics correlates with the last major peak of aridification in the late Miocene (~15 Mya) (Gagnon *et al.* 2019). Interestingly, *Cenostigma* is one of the genera with the most recent diversification (~13.59 Mya) in the Caesalpinia group (56 Mya) (Gagnon *et al.* 2019). Moreover, the diversification of the genus arose between the

Pliocene and Quaternary periods 5.3 Mya ago (Gagnon *et al.* 2019). Although in the Caesalpinia group these ages are recent, when it comes to the rate of evolution of repetitive DNAs, the degree of genomic conservation of *Cenostigma* species is remarkable when compared to other legumes with divergences of >5 Mya (Ribeiro *et al.* 2020).

Genome stability is one of the first traits that can be impaired when two divergent genomes are combined. Genome instability can be considered a hybrid incompatibility phenotype, due to the wide range of sequence classes whose divergence leads to genome instability in the hybrid individual (Dion-Côté and Barbash 2017). The combination of unrelated genomes to form hybrids causes a wide spectrum of genetic and epigenetic changes in the offspring, termed 'genomic shock' (McClintock 1984). However, what happens if the genomes involved in hybrid formation have a high degree of conservation? Comparative analyses among *Cenostigma* species corroborate a scenario of genomic stability that may favour the viability of interspecific hybrids. In this sense, this genomic stability, especially in terms of chromosome number and heterochromatin distribution/composition may provide a better recognition of homoeologous, similarity in epigenetic patterns, leading to stable meiosis and fertility in hybrids (Mason and Batley 2015). Moreover, satDNA are more affected than TEs during the hybridization process, this is congruent with the satDNA differentiation identified here in the *Cenostigma* species (Zagorski *et al.* 2020). In future studies, it will be important to perform similar comparative analysis including such samples identified as potential hybrids with intermediate phenotypes. This will shed light on the potential role of genomic stability in facilitating hybridization in *Cenostigma*.

## CONCLUSION

Advances were reached in the heterochromatin and repetitive fraction characterization of the genomes in seven *Cenostigma* species. Besides the stable chromosome number $2n = 24$, *Cenostigma* is characterized by an intrageneric cytomolecular stability corroborated by three data set obtained here: (i) a conserved CMA+/DAPI$^-$ banding pattern and number of 5S and 35S rDNA sites; (ii) similarity in the composition/abundance of the elements composing the repetitive fraction of the genomes of *C. bracteosum*, *C. laxiflorum*, *C. macrophyllum*, *C. nordestinum*, *C. pluviosum*, and *C. pyramidale*; and (iii) conserved karyotypic distribution of *Tekay* and *Athila* TE elements composing the heterochromatin of the species. The *CemiSat163* satDNA showed few differences in its distribution, thus further studies of the satellitome could yield cytogenetic markers to differentiate the species in the genus. This scenario of genomic stability may be favoured by a similar ecological niche shared by the sampled species that may facilitate the emergence and establishment of hybrids within *Cenostigma* in Caatinga vegetation. Additionally, the relatively recent age of this genus (~13.59 Mya) may be related to an incipient differentiation of the genomes. Finally, the characteristic arboreal habit and relatively long-life cycle of *Cenostigma* species may also be linked to the conservatism of the repetitive fraction in the genus. This is the first infrageneric comparative cytogenomic investigation in the Caesalpinia group, so future studies may reveal whether the heterochromatic stability reported here also occurs in other genera with similar evolutionary/biogeographic histories.

## SUPPLEMENTARY DATA

Supplementary data is available at *Botanical Journal of the Linnean Society* online.

**Table S1.** Genomic composition (expressed in percentage) of individual analysis of repetitive sequences in *C. bracteosum*, *C. laxiflorum*, *C. macrophyllum*, *C. microphyllum*, *C. nordestinum*, *C. pluviosum* and *C. pyramidale* by cluster analysis. Genome size are expressed as Mbp

**Figure S1**. Phylogenetic relationships constructed through FastTree by Neighbor Joining using the reverse transcriptase sequence alignments for Ty3/gypsy-*Athila* A, Ty3/gypsy-*Tekay* B, and the monomer alignment for *CemiSat163* elements satDNA C.

**Figure S2**. MAFFT alignment of the Ty3/gypsy-*Athila* A and Ty3/gypsy-*Tekay* B elements contigs in *C. bracteosum*, *C. laxiflorum*, *C. macrophyllum*, *C. microphyllum*, C. *nordestinum*, *C. pluviosum*, and *C. pyramidale*

**Figure S3.** Mapping of the *CemiSat163* satDNA monomer A and forward and reverse primers B on the alignment of the satDNA *CemiSat163* in *C. bracteosum*, *C. laxiflorum*, *C. macrophyllum*, *C. microphyllum*, C. *nordestinum*, *C. pluviosum*, and *C. pyramidale.*

**Figure S4.** Metaphases of A, *C. pluviosum*; B, *C. macrophyllum*; C, *C. pyramidale*; D, *C. laxiflorum*; E, *C. bracteosum*, and F, *C. nordestinum* showing the Ty3/gypsy-*Tekay* labelled with Cy3-dUTP with pseudocoloured signals in green. The chromosomes are counterstained with DAPI. Scale bar: 10 μm

**Figure S5.** Metaphases of A, *C. microphyllum*; B, *C. pluviosum*; C, *C. macrophyllum*; D, *C. pyramidale*; and E, *C. laxiflorum* showing the Ty3/gypsy-*Athila* labelled with Cy3-dUTP with pseudocoloured signals in red. The chromosomes are counterstained with DAPI. Scale bar: 10 μm.

## ACKNOWLEDGEMENTS

The authors thank the Brazilian agencies Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq number 04/2021) and Fundação de Amparo à Ciência e Tecnologia de Pernambuco (FACEPE-APQ-0970-2.03/15; BIC-1221-2.02/19) for financial support. G.S. thanks to Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) by a post-doctoral grant (Estágio Pós-Doutoral 88881.119479/2016-01). A.M. is financially supported by the Max Planck Society. This study was financed in part by the CAPES (Brazil)—Finance Code 001. G.S. receive productivity fellowship from CNPq (process numbers PQ-312852/2021-5). We also thank the ELIXIR-CZ Research Infrastructure Project (LM2015047) for providing computational resources for RepeatExplorer analysis. We would like to thank Carmelito Vitório da Silva Junior for his assistance in field collections in Catingal (Manoel Vitorino-BA, Brazil).

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## LIMITATIONS OF THE STUDY

The seven *Cenostigma* species analysed in this study are characterized by TE-enriched pericentromeric bands and satellites. However, we were unable to hybridize all elements in all species because of the difficulty in obtaining material for cytogenetic analyses, some material was extremely difficult to germinate and seeds were prone to constant fungal infections. Some species of *Cenostigma* showed a different DNA satellite banding pattern; therefore, it is unclear whether this fraction of the genome evolved similarly in all species of the genus. Extending our comparative cytogenomic approach to the remaining seven species of the genus will help to validate the results observed here.

## DATA AVAILABILITY

All sequencing data used in this study were submitted to the NCBI under the Bioproject no. PRJNA739461. The plastomes presented in this work are made available in Table 1. All other data needed to evaluate the conclusions in the paper are provided in the paper and/or the Supplementary Materials. Additional data related to this study may be requested from the authors.

## REFERENCES

Abadi S, Azouri D, Pupko T *et al.* Model selection may not be a mandatory step for phylogeny reconstruction. *Nature Communications* 2019;**10**:934.

Aecyo P, Marques A, Huettel B *et al.* Plastome evolution in the Caesalpinia group (Leguminosae) and its application in phylogenomics and populations genetics. *Planta* 2021;**254**:1–19.

Alves MAO, Custodio A. Cytogenetics of Leguminosae collected in the State of Ceará. *Revista Brasiliera de Genetica* 1989;**12**:81–92.

Bilinski P, Albert PS, Berg JJ *et al.* Parallel altitudinal clines reveal trends in adaptive evolution of genome size in *Zea mays*. *PLoS Genetics* 2018;**14**:e1007162.

Borges LA, Souza LGR, Guerra M *et al.* Reproductive isolation between diploid and tetraploid cytotypes of Libidibia ferrea (= Caesalpinia ferrea) (Leguminosae): ecological and taxonomic implications. *Plant Systematics and Evolution* 2012;**298**:1371–81.

Brookfield JF. The ecology of the genome—mobile DNA elements and their hosts. *Nature Reviews Genetics* 2005;**6**:128–36.

Costa L, Marques A, Buddenhagen C *et al.* Aiming off the target: recycling target capture sequencing reads for investigating repetitive DNA. *Annals of Botany* 2021;**128**:835–48.

De Jong JH, Fransz P, Zabel P. High resolution FISH in plants–techniques and applications. *Trends in Plant Science* 1999;**4**:258–63.

De Souza TB, Chaluvadi SR, Johnen L *et al.* Analysis of retrotransposon abundance, diversity and distribution in holocentric Eleocharis (Cyperaceae) genomes. *Annals of Botany* 2018;**122**:279–90.

Delgado-Salinas A, Bibler R, Lavin M. Phylogeny of the genus *Phaseolus* (Leguminosae): a recent diversification in an ancient landscape. *Systematic Botany* 2006;**31**:779–91.

Dias Y, Sader MA, Vieira MLC *et al.* Comparative cytogenetic maps of *Passiflora alata* and *P. watsoniana* (Passifloraceae) using BAC-FISH. *Plant Systematics and Evolution* 2020;**306**:51.

Dion-Côté AM, Barbash DA. Beyond speciation genes: an overview of genome stability in evolution and speciation. *Current Opinion in Genetics & Development* 2017;**47**:17–23.

Dodsworth S, Pokorny L, Johnson MG *et al.* Hyb-Seq for flowering plant systematics. *Trends in Plant Science* 2019;**24**:887–91.

Doležel J, Greilhuber J, Suda J. Estimation of nuclear DNA content in plants using flow cytometry. *Nature Protocols* 2007;**2**:2233–44.

Doležel J, Sgorbati S, Lucretti S. Comparison of three DNA fluorochromes for flow cytometric estimation of nuclear DNA content in plants. *Physiologia Plantarum* 1992;**85**:625–31.

DRYFLOR, Banda RK, Delgado-Salinas A, *et al.* Plant diversity patterns in neotropical dry forests and their conservation implications. *Science* 2016;**353**:1383–7.

Fischle W, Wang Y, Jacobs SA *et al.* Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 by Polycomb and HP1 chromodomains. *Genes & Development* 2003;**17**:1870–81.

Gagnon E, Bruneau A, Hughes CE *et al.* A new generic system for the pantropical Caesalpinia group (Leguminosae). *PhytoKeys* 2016;**71**:1–160.

Gagnon E, Ringelberg JJ, Bruneau A *et al.* Global Succulent Biome phylogenetic conservatism across the pantropical Caesalpinia Group (Leguminosae). *New Phytologist* 2019;**222**:1994–2008.

Gao X, Hou Y, Ebina H *et al.* Chromodomains direct integration of retrotransposons to heterochromatin. *Genome Research* 2008;**18**:359–69.

Gerlach WL, Bedbrook JR. Cloning and characterization of ribosomal RNA genes from wheat and barley. *Nucleic Acids Research* 1979;**8**:1869–85.

Glombik M, Bačovský V, Hobza R *et al.* Competition of parental genomes in plant hybrids. *Frontiers in Plant Science* 2020;**11**:200.

González ML, Chiapella JO, Urdampilleta JD. Characterization of some satellite DNA families in *Deschampsia antarctica* (Poaceae). *Polar Biology* 2018;**41**:457–68.

He L, Zhao H, He J *et al.* Extraordinarily conserved chromosomal synteny of Citrus species revealed by chromosome-specific painting. *The Plant Journal* 2020;**103**:2225–35.

Jacobs SA, Khorasanizadeh S. Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. *Science* 2002;**295**:2080–3.

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 2013;**30**:772–80.

Kearse M, Moir R, Wilson A *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 2012;**28**:1647–9.

Kirov I, Khrustaleva L, Van Laere K *et al.* DRAWID: user-friendly java software for chromosome measurements and idiogram drawing. *Comparative Cytogenetics* 2017;**11**:747–57.

Lewis GP. Systematic studies in neotropical '*Caesalpinia* L.' (Leguminosae: Caesalpinioideae), including a revision of the 'Poinchianella-Erythrostemon' group. *Thesis*, University of St Andrews, 1995.

Loureiro J, Rodriguez E, Doležel J *et al.* Two new nuclear isolation buffers for plant DNA flow cytometry: a test with 37 species. *Annals of Botany* 2007;**100**:875–88.

Lower SS, McGurk MP, Clark AG *et al.* Satellite DNA evolution: old ideas, new approaches. *Current Opinion in Genetics & Development* 2018;**49**:70–8.

Lyu H, He Z, Wu CI *et al.* Convergent adaptive evolution in marginal environments: unloading transposable elements as a common strategy among mangrove genomes. *New Phytologist* 2018;**217**:428–38.

Macas J, Novák P, Pellicer J *et al.* In depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe Fabeae. *PLoS One* 2015;**10**:e0143424.

Mallet J. Hybridization as an invasion of the genome. *Trends in Ecology & Evolution* 2005;**20**:229–37.

Marques A, Klemme S, Houben A. Evolution of plant B chromosome enriched sequences. *Genes* 2018;**9**:515.

Marques A, Ribeiro T, Neumann P *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* 2015;**112**:13633–8.

Mascagni F, Giordani T, Ceccarelli M *et al.* Genome-wide analysis of LTR-retrotransposon diversity and its impact on the evolution of the genus Helianthus (L.). *BMC Genomics* 2017;**18**:1–16.

Mason AS, Batley J. Creating new interspecific hybrid and polyploid crops. *Trends in Biotechnology* 2015;**33**:436–41.

Mata-Sucre Y, Costa L, Gagnon E *et al.* Revisiting the cytomolecular evolution of the *Caesalpinia* group (Leguminosae): a broad sampling reveals new correlations between cytogenetic and environmental variables. *Plant Systematics and Evolution* 2020a;**306**:1–13.

Mata-Sucre Y, Sader M, Van-Lume B *et al*. How diverse is heterochromatin in the *Caesalpinia* group? Cytogenomic characterization of *Erythrostemon hughesii* Gagnon & G.P. Lewis (Leguminosae: Caesalpinioideae). *Planta* 2020b;**252**:1–14.

McCann J, Macas J, Novák P *et al*. Differential genome size and repetitive DNA evolution in diploid species of *Melampodium* sect. *Melampodium* (Asteraceae). *Frontiers in Plant Science* 2020;**11**:362.

McClintock B. The significance of responses of the genome to challenge. *Science* 1984;**226**:792–801.

Min J, Zhang Y, Xu RM. Structural basis for specific binding of Polycomb chromodomain to histone H3 methylated at Lys 27. *Genes & Development* 2003;**17**:1823–8.

Negi P, Rai AN, Suprasanna P. Moving through the stressed genome: emerging regulatory roles for transposons in plant stress response. *Frontiers in Plant Science* 2016;**7**:1448.

Neumann P, Novák P, Hoštáková N *et al*. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mobile DNA* 2019;**10**:1–17.

Nguyen TH, Waminal NE, Lee DS *et al*. Comparative triple-color FISH mapping in eleven *Senna* species using rDNA and telomeric repeat probes. *Horticulture, Environment and Biotechnology* 2021;**62**:927–35.

Nielsen PR, Nietlispach D, Mott HR *et al*. Structure of the HP1 chromodomain bound to histone H3 methylated at lysine 9. *Nature* 2002;**416**:103–7.

Novák P, Ávila Robledillo L, Koblížková A *et al*. TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Research* 2017;**45**:e111–e111.

Novák P, Neumann P, Macas J. Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. *Nature Protocols* 2020;**15**:3745–76.

Novák P, Neumann P, Pech J *et al*. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics* 2013;**29**:792–3.

Oksanen J, Simpson GL, Blanchet FG *et al*. Package 'vegan'. *Community ecology package* 2013;**2**:1–295.

Oliveira MAS, Nunes T, Dos Santos MA *et al*. High-throughput genomic data reveal complex phylogenetic relationships in *Stylosanthes* Sw (Leguminosae). *Frontiers in Genetics* 2021;**12**:727314.

Parisod C, Alix K, Just J *et al*. Impact of transposable elements on the organization and function of allopolyploid genomes. *New Phytologist* 2010;**186**:37–45.

Pedrosa A, Sandal N, Stougaard J *et al*. Chromosomal map of the model legume lotus *Japonicus*. *Genetics* 2002;**161**:1661–72.

Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution* 2009;**26**:1641–50.

Raskina O, Barber JC, Nevo E *et al*. Repetitive DNA and chromosomal rearrangements: speciation-related events in plant genomes. *Cytogenetic and Genome Research* 2008;**120**:351–7.

Raskina O, Belyayev A, Nevo E. Quantum speciation in *Aegilops*: Molecular cytogenetic evidence from rDNA cluster variability in natural populations. *Proceedings of the National Academy of Sciences of the United States of America* 2004;**101**:14818–23.

Ribeiro T, Vasconcelos E, dos Santos KGB *et al*. Diversity of repetitive sequences within compact genomes of *Phaseolus* L. beans and allied genera *Cajanus* L. and *Vigna Savi*. *Chromosome Research* 2020;**28**:139–53.

Roa F, Guerra M. Distribution of 45S rDNA sites in chromosomes of plants: structural and evolutionary implications. *BMC Evolutionary Biology* 2012;**12**:225.

Rodrigues PS, Souza MM, Melo CAF *et al*. Karyotype diversity and 2C DNA content in species of the *Caesalpinia* group. *BMC Genetics* 2018;**19**:1–10.

Schley RJ, Pellicer J, Ge XJ *et al*. The ecology of palm genomes: repeat-associated genome size expansion is constrained by aridity. *New Phytologist* 2022;**236**:433–46.

Shannon CE. A mathematical theory of communication. *The Bell System Technical Journal* 1948;**27**:623–56.

Simpson EH. Measurement of diversity. *Nature* 1949;**163**:688–688.

Souza G, Costa L, Guignard MS *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* 2019;**38**:13–23.

Team R Core. R: a language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing, 2019.

Topalian J, González ML, Chiapella JO *et al*. Retrotransposons in *Deschampsia antarctica* E. Desv.(Poaceae) genome: diversity, abundance and chromosomal distribution. *Polar Science* 2022;**31**:100762.

Usai G, Mascagni F, Vangelisti A *et al*. Interspecific hybridisation and LTR-retrotransposon mobilisation-related structural variation in plants: A case study. *Genomics* 2020;**112**:1611–21.

Vaio M, Nascimento J, Mendes S *et al*. Multiple karyotype changes distinguish two closely related species of *Oxalis* (*O. psoraleoides* and *O. rhombeo-ovata*) and suggest an artificial grouping of section Polymorphae (Oxalidaceae). *Botanical Journal of the Linnean Society* 2018;**188**:269–80.

Van-Lume B, Esposito T, Diniz-Filho JAF *et al*. Heterochromatic and cytomolecular diversification in the *Caesalpinia* group (Leguminosae): relationships between phylogenetic and cytogeographical data. *Perspectives in Plant Ecology, Evolution and Systematics* 2017;**29**:51–63.

Van-Lume B, Mata-Sucre Y, Báez M *et al*. Evolutionary convergence or homology? Comparative cytogenomics of *Caesalpinia* group species (Leguminosae) reveals diversification in the pericentromeric heterochromatic composition. *Planta* 2019;**250**:2173–86.

Venner S, Feschotte C, Biemont C. Dynamics of transposable elements: towards a community ecology of the genome. *Trends in Genetics* 2009;**25**:317–23.

Zagorski D, Hartmann M, Bertrand YJK *et al*. Characterization and dynamics of Repeatomes in closely related species of Hieracium (Asteraceae) and their synthetic and apomictic hybrids. *Frontiers in Plant Science* 2020;**11**:591053.

Zizka A, Silvestro D, Andermann T *et al*. CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution* 2019;**10**:744–51.