Title: Nuclear genome profiling of two Mexican orchids of the genus *Epidendrum*

Miguel A. Alcalá-Gaxiola[1,2], Gerardo A. Salazar[1], Eric Hágsater[3], Mónica A. Flores-Iniestres[1,2], Lidia I. Cabrera[1], Aldo I. Aviña-Rivera[1], Pedro Mercado-Ruaro[1], Susana Magallon[1], Carolina Granados Mendoza[1], Aleida Núñez-Ruiz[4], Gloria Soldevila[4,5], Araxi Urrutia[6], Rubi N. Meza-Lázaro[1,3]

[1]Instituto de Biología, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico.

[2]Posgrado en Ciencias Biológicas, Universidad Nacional Autónoma de México, Av. Universidad 3000, Coyoacán, Mexico City 04510, Mexico.

[3]Herbario AMO, Lomas de Chapultepec, Miguel Hidalgo, Mexico City 11000, Mexico.

[4]Laboratorio Nacional de Citometría de Flujo, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico.

[5]Departamento de Inmunología, Instituto de Investigaciones Biomédicas, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico.

[6]Instituto de Ecología, Universidad Nacional Autónoma de México, Mexico City 04510, Mexico.

Keywords: genome size, heterozygosity, k-mer, Orchidaceae, ploidy level, repeatome

ABSTRACT

Characterizing genomic properties such as genome size, ploidy level, heterozygosity, and repetitive DNA proportion and composition without relying on genome assembly is crucial for profiling the genomes of non-model species. Little is known about the nuclear genome of the large neotropical orchid genus *Epidendrum*. This study compares genome profiles of *Epidendrum anisatum* and *E. marmoratum,* using flow cytometry and k-mer analysis

approaches, as well as bioinformatics ploidy level estimation and repeatome characterization. Multiple depths of coverage, *k* values, and k-mer-based tools for genome size estimation were explored. A 2.3-fold genome size difference was found between the two species, ranging from 2.35 to 2.80 Gb ($\bar{x}$ = 2.58 Gb) for *E. anisatum* and from 0.99 to 1.24 Gb ($\bar{x}$ = 1.11 Gb) for *E. marmoratum*. Both species were identified as diploid with no evidence of strict partial endoreplication. The most important aspects to be taken into account to improve genome size estimation were heterozygosity, depth of coverage, and the maximum k-mer coverage. The genomes of both species were found to be highly repetitive (63–73%) and heavily dominated by Ty3-gypsy retrotransposons, particularly those of the Ogre family. Additionally, the genome of *E. anisatum* was characterized by the presence of a 172 bp satellite (*AniS1*), which represented 11% of the genome size. Together, both Ty3-gypsy transposons and *AniS1* shape the genome size difference between the two *Epidendrum* genomes. This study provides the first genome profiling for species in the genus *Epidendrum*, but also highlights the importance of using flow cytometry, cytogenetic approaches and bioinformatics techniques in combination for genome profiling.

INTRODUCTION

Plant genomes exhibit a particularly large variation in their size, ploidy levels and repetitive DNA content (Wang *et al*. 2021). These characteristics play important roles in plant lineage evolution (Li *et al*. 2017) and are important considerations prior to whole genome sequencing and genome assembly efforts (Dominguez Del Angel *et al*. 2018). The characterization of genome size, heterozygosity, and genome repetitiveness, also referred to as genome profiling (Jenike *et al.* 2025), does not require genome assembly –a labor-intensive and complex process.

Genome size is usually estimated using laboratory techniques on living cells, specifically Feulgen densitometry and flow cytometry (Hardie *et al.* 2002; Leitch *et al.* 2019). These methods rely on staining techniques to quantify total nuclear DNA from live tissues, which, in many cases, are not available. Flow cytometry is the gold standard for genome size estimation, but whole-genome endoreplication (also known as conventional endoreplication; CE) and strict partial endoreplication (SPE) can confound this method. CE and SPE are cell cycle programs during which cells replicate their genomes without division (Brown *et al*. 2017; Shu *et al*. 2018). The first process results in polyploid cells (Shu *et al*. 2018), while the second results in cells that replicate a fixed (F) fraction of the genome (Brown *et al.* 2017).

Rapid bioinformatics methods, independent of whole genome assembly, are available for genome profiling, opening the possibility of exploring genome evolution in highly diverse non-model groups with no reference genome available, such as orchids, utilizing specimens from biological collections and museums. Bioinformatics techniques generally use genomic sequencing data to calculate the average depth of coverage in the data set to estimate nuclear genome size, utilizing the principles of the Lander-Waterman equation (Lander and Waterman 1988; Pflug *et al.* 2020). These methods can be broadly divided into two types: read mapping to low-copy loci, which relies on a high-quality reference genome to identify

low-copy loci (Pflug *et al*. 2020), and k-mer analysis, which is a reference genome free technique for genome profiling.

K-mer analysis relies on plotting the number of times a substring of a fixed length k (k-mer) is found within the dataset (k-mer coverage), against the k-mer coverage frequency, to statistically estimate the genome size and other properties such as heterozygosity, sequencing error rate, and repetitive proportion. However, k-mer analysis is often distrusted for estimating nuclear genome size because of discrepancies with laboratory estimations (Hesse 2023). These discrepancies can result from several factors that can influence the final genome size estimation obtained from k-mer analysis. These factors can be methodological issues, such as the choice of the value of k (k value) during k-mer counting and the techniques used to estimate genome size from the resulting histogram. Additionally, there are intrinsic challenges related to the data and the genomic characteristics of the studied organisms, including high sequencing error rates, significant heterozygosity, the occurrence of polyploidy, and the amount of repetitive DNA in the genome (Pflug *et al.* 2020).

The abundance of repetitive DNA is the main factor behind the evolutionary changes in genome size across various eukaryotic lineages (Macas *et al.* 2015; Pellicer *et al.* 2018; Hloušková *et al.* 2019), alongside, to a lesser extent, polyploidization (e.g. Cordeiro *et al.* 2022). Some of the first methods for studying the repetitive DNA fraction of the genome relied on the reassociation kinetics of DNA using C_0t curves (Britten and Kohne 1968). C_0t curves are a biochemical approach that allows for the simultaneous estimation of genome size and the relative frequencies of repetitive DNA (Palmer and Black IV 1997). Massively parallel sequencing has introduced additional methods for DNA fraction estimation and repeat type identification. Repeat identification methods can be classified into *de novo* methods, methods relying on the comparison with sequence and structure databases, and hybrids of the previous two (Liao *et al.* 2023).

Repetitive DNA is typically classified into two types: tandem repeats, known as satellites, and interspersed repeats, referred to as transposons (Lee and Kim 2014). Transposons can then be classified into class I and class II transposons, based on whether an RNA intermediary is present or not, respectively. Long terminal repeat retrotransposons (LTRs) are the most common type of class I transposons in plants, particularly those belonging to the Ty1-copia and Ty3-gypsy superfamilies (Neumann *et al*. 2019).

The neotropical orchid genus *Epidendrum* (subtribe Laeliinae) has previously been proposed as a promising model system for evolutionary and ecological studies (e.g. Pinheiro and Cozzolino 2013; Granados Mendoza *et al.* 2020; Pessoa *et al.* 2021), both because of its outstanding taxic diversity (over 1800 species; Karremans 2021; Moonlight et al. 2024) and morphological and ecological disparities (Hágsater and Soto 2005). Despite good morphological and ecological descriptions, little is known about their nuclear genome characteristics. This study aims to use two closely related endemic Mexican species, *E. anisatum* and *E. marmoratum*, to provide the first genomic profiling for this genus focusing on the following aspects: 1) their genome sizes using both flow cytometry and bioinformatics methods; 2) different k-mer analysis approaches for estimating genome size; 3) bioinformatics ploidy level estimation and chromosome counting; and 4) the repeat composition, proportion, and abundance.

METHODS

Genome size estimation by flow cytometry

Leaf tissue samples of *Epidendrum anisatum* (voucher: *Salazar et al. 7370*) and *E. marmoratum* (*Salazar 10329*), and ovary tissue and pollinia (suitable for identifying the 2C and 1C nuclei population; Trávníček *et al*. 2015) were processed to estimate genome size by flow cytometry at the Laboratorio Nacional de Citometría de Flujo (LabNalCit;

https://labnalcit.org/). The genome size was estimated following Doležel and Bartoš (2005). Nuclei were isolated in LB01 buffer (15 mM Tris, 2 mM EDTA, 20 mM NaCl, 80 mM KCl, 0.5 mM spermine tetrahydrochloride, 0.1% v/v Triton 300, and 15 mM β-mercaptoethanol, pH 7.5; Doležel 1989). The suspension was filtered through a 30–50 µm pore CellTrics™ filter (SYSMEX; https://www.sysmex.com) and treated with 50µg/mL RNAse A (Qiagen) and 50µg/mL propidium iodide (Sigma-Aldrich). Young leaves of *Pisum sativum* (4.45 pg/1C; Doležel *et al.* 1998) and peripheral blood mononuclear cells (PBMCs) from healthy individuals were used as internal standard references. Samples were analyzed in a CytoFLEX S (Beckman-Coulter), individually and in combination with the internal reference(s). The analysis of cytometry data was performed using FlowJo® v. 10 (https://www.flowjo.com/). A genome size value for each species was calculated as the average of the minimum and maximum 1C/2C values obtained from three independent analyses of the DNA content histograms.

DNA sequencing

For library preparation, total genomic DNA was extracted from the leaf tissue of the same *Epidendrum* specimens noted earlier using the Qiagen DNeasy Plant Mini Kit. Twelve lysates of each sample were sequentially loaded onto a single DNeasy Mini spin column and centrifuged to obtain approximately 8 µg of DNA per sample. The quality and quantity of the genomic DNA were assessed by agarose gel electrophoresis and a NanoDrop 2000/2000c spectrophotometer (Thermo Fisher Scientific). DNA extractions were sent to Macrogen Inc. (Seoul, South Korea) for shotgun library preparation and sequencing. Libraries were prepared by random fragmentation and adapter ligation, and 150 bp paired-end sequencing was carried out in an Illumina HiSeq 2500 sequencer (Illumina, Inc., San Diego, California, US).

The sequences were trimmed and filtered using Trimmomatic 0.3.9 (Bolger *et al.* 2014) with the following parameters: ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10:2:True

LEADING:3 TRAILING:3 MINLEN:100. We assessed the quality of the trimmed data using FastQC 0.11.9 (Andrews 2010).

Ploidy level estimation and chromosome counts

The ploidy level for each species was estimated using the distribution of bi-allele frequencies through nQuire (Weiß *et al*. 2018; https://github.com/clwgg/nQuire), which implements a Gaussian Mixture Model to test three possible ploidy models: diploidy, triploidy, and tetraploidy. To estimate the allele frequencies, the filtered sequences were mapped to a reference of low-copy nuclear genes for each species (Granados Mendoza *et al*. 2020) using BWA-MEM 2.2.1 (Li 2013) and SAMtools 1.15.1 (Danecek *et al*. 2021). The bi-allelic variants were filtered using the *denoise* command implemented in nQuire, and the ploidy level was assessed using the likelihood differences obtained from the *lrdmodel* command.

Chromosome number was counted from developing root tips, which were pretreated using 0.002 M 8-hydroxyquinoline and fixed in absolute ethanol:glacial acetic acid (3:1). To prepare the slides, the root tips were hydrolyzed in 1 M HCl for 12 min at 60°, stained for one hour with Feulgen dye, and squashed carefully to disperse the cells. For the preparation of permanent slides, the liquid nitrogen technique was used; dispersed chromosomes were counted in intact cells.

Genome size estimation by k-mer analysis

To assess the suitability of the whole dataset and estimate the minimum coverage required for genome size estimation, the depth of coverage of both datasets was calculated based on the flow cytometry genome size values. Subsequently, the datasets were subsampled to simulate depths of coverage of 5×, 10×, 20×, 30×, and 40×. The depth of coverage of the datasets (both whole and subsampled) should be sufficient to resolve the error peak from the haploid peak (Ranallo-Benavidez 2020). Canonical k-mer counting and k-mer frequency plotting were performed using Jellyfish v.1.3 (Marçais and Kingsford 2011), with an exploratory

maximum k-mer coverage value of 10,000 and subsequently a value of two million (Hesse 2023). To encompass a k-mer range other than the most common options (e.g. 21, 31), the k-values = 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 41, 43, and 45 were used. Values are required to be odd to avoid palindromic k-mers. Genome size estimations were then carried out for both the whole datasets and each of the subsampled datasets.

Genome size for each k-value was estimated using three different software tools: CovEst 0.5.6 (Hozza *et al*. 2015), GenomeScope (Vurture *et al.* 2017), and findGSE 0.1.0 (Sun *et al*. 2018). Each uses a different statistical approach to calculate the depth of coverage, applying the principles of the Lander-Waterman equation. CovEst uses truncated Poisson distributions within a probabilistic framework, GenomeScope performs a non-linear minimum square optimization to adjust a mix of negative binomial distributions, and findGSE employs a mixed model, adjusting the k-mer frequencies iteratively using a skewed normal distribution. CovEst implements the “basic” and “repeats” algorithms, and findGSE implements the “hom” and “het” algorithms, which allow for examining the effects of high repetitive DNA proportion and heterozygosity, respectively. Heterozygosity was also estimated using GenomeScope and k = 21. Plots were generated using R 4.4.2 (R Core Team 2024) and the ggplot2 package (Wickham 2016).

Repetitive DNA proportion and abundance estimation

The repeatomes of *Epidendrum anisatum* and *E. marmoratum* were characterized using the RepeatExplorer2 pipeline (Novák *et al*. 2020) within the graphical interface of the Galaxy platform (https://repeatexplorer-elixir.cerit-sc.cz/) and the default parameters. RepeatExplorer2 can be classified as a hybrid software for repeat detection (Liao *et al*. 2023), since it performs homology-based annotation, utilizing the RexDB databases (Neumann *et al*. 2019) and *de novo* detection through graph clustering (*seqclust* program) and TAREAN

(Novák *et al*. 2017). TAREAN calculates the probability of a satellite using the C index (connected component index) and P index (pair completeness index), and its proportion based on the number of reads. The C index is a measure for identifying clusters representing tandem repeats (a value of one represents an exact tandem repeat). The P index is a measure of the proportion of unbroken read pairs (Novák *et al*. 2017).

Samples were analyzed using the individual and comparative protocols (Novák *et al.* 2020). The individual protocol allows for a detailed analysis of repetitive DNA and estimation of non-repetitive DNA (singlets) for a single species. The comparative protocol identifies shared clusters, allowing for comparisons by cluster instead of solely by repeat type. It does this by merging the reads from both samples, proceeding to the clustering and annotation steps, and finally separating the clusters by species. The automatic annotation step was performed using the RexDB Viridiplantae 3.0 database, and the manual annotation step was based on the cluster's graph shape analysis and information from the automatic annotation.

RESULTS

Genome size estimation by flow cytometry

Fluorescence intensity of DNA staining of the nuclei revealed a peak for *Pisum sativum* leaves (internal standard reference) and a series of distinct peaks for *E. anisatum* and *E. marmoratum* derived from the ovary, pollinium, and leaf tissues (Fig. 1). Histograms showed multiple peak positions with no interpeak values, mirroring the profiles expected from a conventional endoreplication process. The pollinium histograms showed peaks corresponding to 1C and 2C, while the ovary tissue showed peaks at 2C and 4C. For *E. marmoratum*, the histograms indicated an additional 8C peak in ovary tissue and a 4C peak in pollinia (Fig. 1). The 1C value in gigabases (Gb; calculated from mass in pg) of *E. anisatum* ranged from 2.35 to 2.80 Gb ($\bar{x}$ = 2.58 Gb) and that of *E. marmoratum* from 0.99 to 1.24 Gb ($\bar{x}$ = 1.11 Gb).

Genome size estimation by k-mer analysis and ploidy estimation

We obtained 760.0 million reads for *E. anisatum* and 765.3 million reads for *E. marmoratum* after trimming and filtering. We found both species to be diploids because the diploid model Δlog-likelihood was the lowest when compared with the free model likelihood (Table 1). The estimated maximum depth of coverage of the entire data set reached over 40× for *E. anisatum* and over 100× for *E. marmoratum.* The k-mer distribution histograms indicated that the error, heterozygous, and homozygous peaks can be clearly distinguished when using the whole datasets (Fig. 2). However, the error peaks were indistinguishable from the homozygous peak when analyzing low-depth of coverage datasets (5×). At higher coverage depths (greater than 10×), these two peaks became distinct, particularly in *E. marmoratum*. The homozygous and heterozygous peaks could be distinguished for some k-values at a 20× depth of coverage in *E. anisatum* (k-value = 17–21) and at a 30× depth of coverage in *E. marmoratum*. The error, the heterozygous, and the homozygous peaks could be clearly identified in the histograms for all k-values of *E. anisatum* starting at a 30× depth of coverage and over, and in the histograms of *E. marmoratum* at a 40× depth of coverage and over (Supplementary Data Figure S1).

The full-sampling histograms (Fig. 2) illustrate the influence of the k-value on peak height and position. The heterozygous and homozygous peaks moved toward lower k-mer coverage (left), and the frequency of k-mers increases with higher k-values. In *E. anisatum,* the heterozygous peak was higher than the homozygous peak for every value of k (Fig. 2). This is mirrored by the heterozygosity estimation by GenomeScope, which was likewise higher in *E. anisatum* (1.7%) than in *E. marmoratum* (0.46%).

All of the tools estimated a higher genome size for *E. anisatum* than for *E. marmoratum*. The comparison of tools for genome size estimation yielded a general pattern in which genome size increases with the k-value (Fig. 3), mirroring the behavior of the homozygous peaks in the k-mer distribution histograms. FindGSE-het results (*E. anisatum*:

1.46–2.65 Gbp; *E. marmoratum*: 0.7–1.15 Gbp) were within the range of flow cytometry genome size estimation for both species, except at low k-values. In contrast, findGSE-hom substantially overestimated genome size for *E. anisatum* (2.85–5.3 Gbp), while being more consistent with the flow cytometry estimation for *E. marmoratum* (0.8–1.3 Gbp). GenomeScope's estimate was within the range of *E. marmoratum* (0.69–1.06 Gbp) and slightly underestimated the genome size for *E. anisatum* (0.71–2.21 Gbp). CovEst-basic tended to underestimate the genome size in both species (*E. anisatum*: 0.3–1.86 Gbp; *E. marmoratum*: 0.24–1.04 Gbp). Lastly, CovEst-repeats did not show a consistent increase in genome size with increasing k-values, instead exhibiting strong fluctuations (*E. anisatum*: 0.53–4.11 Gbp; *E. marmoratum*: 0.33–1.95 Gbp). The estimated genome sizes using a maximum count value of 10,000 were lower for all tools in both species compared to using a maximum count value of 2 million. Only findGSE-het and CovEst-repeats yielded values within the range of genome size estimated by flow cytometry for some k-values in *E. anisatum*, as did findGSE-hom and CovEst-repeats for *E. marmoratum* (Supplementary Data Figure S2).

The analysis of subsampled data simulating different depths of coverage yielded consistent genome size estimations similar to those of the whole datasets at depths of coverage ≥ 20×. Particularly, findGSE-het and GenomeScope results are almost always within the estimation range of flow cytometry. CovEst-basic and CovEst-repeats follow the same pattern as with the whole dataset estimations for both species in all samplings. CovEst-basic underestimates genome size, and CovEst-repeats yields results that strongly fluctuate with k value. The exception is for *E. marmoratum* at a 10× depth of coverage, where the CovEst-repeats estimations always fall within the genome size estimation range of flow cytometry, and the fluctuations are less pronounced (Supplementary Data Figure S3).

Estimation of repetitive DNA proportion and abundance

To estimate the proportion of repetitive DNA, the individual protocol analyzed reads corresponding to depths of coverage of 0.06× for *Epidendrum anisatum* and 0.43× for *E. marmoratum*. According to this protocol, both species have highly repetitive genomes, with 73% and 63% repetitive content, respectively (Table 2). These proportions corresponded to 1.72–2.04 Gb of repetitive DNA in *E. anisatum* and 0.62–0.78 Gb of repetitive DNA in *E. marmoratum*, according to the flow cytometry estimation (Fig. 4).

The dominant repeat type in both genomes was the Ty3-gypsy Ogre family, an LTR retrotransposon group that occupies 33.37% of the *E. anisatum* genome and 32.1% of the *E. marmoratum* genome. The most notable difference between the two species was in the abundance of a single satellite (henceforth *AniS1*), a 172 bp monomer with 44% GC composition that occupies more than 10% of the genome of *E. anisatum*, but is barely present in *E. marmoratum* (11.5% in *E. anisatum* and <0.01% in *E. marmoratum*). The satellite cluster has a C index (connected component index) score of 0.982 and a P index (pair completeness index) score of 0.485.

The genomes of the two *Epidendrum* species analyzed also differ in the proportions of unidentified repeats (2.85% in *E. anisatum* and 7.99% in *E. marmoratum*, differing in 5.14 percentage points) and unidentified Class I transposons (5.33% in *E. anisatum* and 0.43% in *E. marmoratum*, differing in 4.19 percentage points). Other repetitive DNA proportions show a difference of less than three percentage points between the two species.

The comparative protocol showed a 67% total repeat proportion, which falls between the estimated repeat proportions of the two species according to the results of the individual protocol. However, the repeat proportions obtained by the comparative protocol for *E. anisatum* and *E. marmoratum* (76% and 57%, respectively) differed from those obtained by the individual protocol. The comparative protocol annotated 170 clusters, of which 19 were not shared between the two species: 17 were unique to *E. anisatum,* and two were unique to

E. marmoratum. Additionally, in 16 of the shared clusters, one of the species dominated with at least 95% of the reads in the clusters. In 15 of these, *E. anisatum* dominated, while *E. marmoratum* only dominated in one. The comparative protocol agreed with the results of the individual protocol in terms of the repeat composition of both species. The most abundant cluster was the same satellite cluster as in the individual protocol, which is shared but highly dominated by *E. anisatum*, accounting for more than 99% of the reads in the cluster. Similarly, the dominant repeat type in both species was the Ogre family of retrotransposons. A cluster-by-cluster comparison of the number of reads in both species using the comparative protocol showed that most Ogre clusters have similar proportions between the two species (Fig. 5).

DISCUSSION

Genome size estimation using flow cytometry.

Genome endoreplication is a component of cell differentiation processes commonly observed in plants (Shu *et al.* 2018). In contrast, SPE occurs mainly in orchids (Brown *et al*. 2017). Conventional and partial endoreplication can occur in multiple rounds, producing nuclear populations of 4C, 8C, 16C, and so on, and 4E, 8E, 16E, and so on, respectively (Brown *et al*. 2017; Loureiro *et al*. 2023). Endoreplicated nuclei can be more frequent than the 2C nuclei in some tissues (e.g., leaves and stems), which can be absent or undetected by the cytometer.

Ovary tissue and pollinia are less prone to endoreplication and are suitable for identifying the 2C nuclei population and estimating genome size (Trávníček *et al*. 2015). For example, although *E. anisatum* and *E. marmoratum* DNA content histograms have the same peaks in leaves and ovaries, the 2C peak is much lower than the 4C peak in the leaf histograms (Fig. 1). Such small peaks may be overlooked, and the genome size could be overestimated. The first peak in the pollinia histograms, corresponding to the reduced 1C gamete value, may cross-validate the genome size estimated from other tissues (Fig. 1). In

Epidendrum, the 1C genome size estimation (*E. anisatum*: 2.41–2.81 Gb; *E. marmoratum* 1.11–1.29 Gb) resulted in half the 2C genome size calculated from both the ovary and leaf tissues (2.44–2.85 Gb; 1.05–1.22 Gb, respectively), suggesting that the first peak in the ovary and leaf was correctly identified and does not represent a round of endoreplication.

The ovary and leaf histograms did not present prominent intermediary peaks indicative of SPE, which agrees with a previous study in which no SPE was found to occur in Laeliinae (Brown *et al*. 2017). In contrast, subtribe Pleurothallidinae, the sister group of Laeliinae, includes clades with evidence of SPE (Brown *et al*. 2017; Chumová *et al*. 2021). Difficulties for cytometric estimation of genome size can thus be taxon-specific. Therefore, cross-validating flow cytometry and bioinformatics results can be the most effective method for estimating plant genome size, especially when only tissues suspected to show significant endoreplication, such as leaves, are available.

Genome size estimation by k-mer analysis

Genome size estimation based on k-mer analysis is a powerful approach, but its accuracy can be affected by factors intrinsic to the genome, such as repetitive DNA proportion, heterozygosity, ploidy level, and genome size itself, as well as extrinsic factors associated with a particular dataset, such as sequencing error and dataset size. The choices of tools and parameters made during analysis can also deeply affect the genome size estimation and are closely related to the above mentioned intrinsic and extrinsic factors. The genomes of *Epidendrum anisatum* and *E. marmoratum* differ in terms of size, heterozygosity, and proportion of repetitive DNA, enabling us to assess how the choice of tools and their parameters influences the estimation of genome size, identify the main pitfalls, and improve the estimation.

A highly repetitive genome, such as that of *E. anisatum*, tends to have k-mers with high coverage values, representing the highly repeated parts of the genome. Selecting an

appropriate maximum k-mer coverage value during the k-mer counting steps is the primary parameter for accurately estimating genome size. This parameter is genome-specific and tends to increase with the genome's repetitive proportion (Hesse 2023). Although it underestimated genome sizes, the exploratory maximum k-mer coverage value of 10,000 provided a glimpse into the size, repetitiveness, and heterozygosity of the *Epidendrum* genomes analyzed. Increasing the maximum k-mer coverage to 2 million notably improved genome size estimation for both species. However, *E. anisatum* probably requires an even higher increase in this value to include the most highly repetitive k-mers in the histogram, as recommended by Hesse (2023).

To model repeats and sequencing errors in genome size estimation, CovEst provides the Model RE (Hozza *et al*. 2015) to obtain reasonable estimates of genome sizes from data at 1× genome coverage or less, using a k-value of 21. However, our k-value and simulated coverage using CovEst-repeats were only consistent and accurate at 10× depth of coverage for *E. marmoratum*, which may be evidence of the algorithm's robustness when applied to real datasets with low depth of coverage and low heterozygosity.

The heterozygosity levels of the two species analyzed resulted in three distinct peaks; the first is the sequencing error peak, the second is the heterozygous peak, which occurs at exactly half the k-mer coverage value of the third peak, the homozygous peak (Hozza *et al*. 2015; Vurture *et al*. 2017; Sun *et al*. 2018). Distinguishing between error, heterozygous, and homozygous peaks is crucial for assessing the adequacy of coverage for genome estimation. The sequencing error peak was clearly observed in the histograms of *E. anisatum* at a depth of coverage of 20× or higher, and in *E. marmoratum* at a depth of coverage of 10× or higher, likely due to the smaller genome of *E. marmoratum*. However, a depth of coverage of 20×–25× or higher, generally recommended for diploid genomes (Vurture *et al*. 2017; Hesse 2023), performed better, allowing for the identification of heterozygous and homozygous

peaks and enabling users to establish the parameters necessary for tools like findGSE-het to offset the impact of heterozygosity on the final genome size estimation. In contrast, when heterozygosity was not taken into account, findGSE-hom overestimated genome size, particularly in the highly heterozygous genome of *E. anisatum*.

The k-value of 21 is the most frequently suggested for genome size estimation (Hozza *et al*. 2015; Vurture *et al*. 2017; Pflug *et al*. 2020). Changing the k-value has been recommended for highly repetitive genomes (Vurture *et al*. 2017). This assertion is supported by the 10,000 maximum k-mer coverage experiment, which produces results closer to those of flow cytometry using higher k-values. In contrast, the two-million maximum k-mer coverage value estimations showed that k-values ≥ 17 produce very consistent genome size estimations with most tools (notable exceptions are Cov-Est repeats and findGSE-hom), emphasizing the importance of the maximum k-mer coverage value parameter.

K-mer analysis is a crucial tool for studying the genomes of non-model organisms that have not yet been fully assembled. For diploid organisms like *E. anisatum* and *E. marmoratum*, tools such as GenomeScope and findGSE-het, using k-values of 17 or higher, consistently provide estimates that align well with flow cytometry results. Learning about the complex genomic characteristics of a particular species through standard tools using default parameters can create a positive feedback loop that improves the choice of the tools and parameters for genome size estimation.

Repetitive DNA composition in *Epidendrum anisatum* and *E. marmoratum*

Epidendrum anisatum and *E. marmoratum* have relatively small genomes, similar to those of most angiosperms, which typically have genomes smaller than 5 Gb/1C (Dodsworth *et al.* 2015). Repetitive DNA is considered the main driver of genome size variation in orchid nuclear genomes (Mehrotra and Goyal 2014), which span a 251-fold interval (from approximately 0.22 pg to 55.4 pg; Leitch *et al.* 2009; Chumová *et al.* 2021).

Repetitive DNA sequences are present in all higher plants and can account for up to 90% of the genome size in some species (Mehrotra and Goyal 2014). For example, previously reported repetitive DNA proportions in orchid genomes range from 9.94% in *Lepanthes clareae* (Chumová *et al.* 2021) to 88.9% in *Cymbidium goeringii* (Chung *et al*. 2022). The LTR transposon superfamilies Ty1-copia and Ty3-gypsy are the most abundant repetitive DNA types in a wide range of plant species (Neumann *et al*. 2019). Compared to other plant groups, LTR transposons are highly abundant in monocotyledons, accounting for the majority of the genome (Kejnovsky *et al*. 2012) and have a very high turnover rate, with both gains and losses of transposable element (TE) sequences (*i.e.,* in grasses: Vicient *et al*. 2001; Giraud *et al*. 2021; in *Oryza*: Ma *et al*. 2004; Piegu *et al*. 2006; Vitte *et al*. 2007; in Pleurothallidinae orchids: Chumová *et al*. 2021). Ty1-copia transposons are very scarce (<1%) in *E. anisatum* and *E. marmoratum* compared to other monocotyledons, such as ~20% in *Zea mays* and ~30% in *Oryza sativa* (Haberer *et al*. 2005). In contrast, Ty3-gypsy transposons are very abundant in the two *Epidendrum* species, particularly those of the Ogre family, a group of exceptionally large, plant specific retrotransposons which can reach up to 25 kb in length (Macas and Neumann 2007). Similarly, the species with the lower Ty1-copia percentages in the subtribe Pleurothallidinae have high Ty3-gypsy percentages (i.e. *Anathallis sanchezii* 0.24% *copia*, 48.3% *gypsy*; *Platystele misasiana* 0.02% *copia*, 59.46% *gypsy*; Chumová *et al*. 2021). This suggests that, in *Epidendrum* and Pleurothallidinae, Ty3-gypsy elements have a relatively higher rate of success in terms of amplification and subsequent retention in the genome compared to Ty1-copia elements (Kejnovsky *et al*. 2012).

Multiple cases of extensive chromosomal variation have been reported in *Epidendrum*, involving recent and ancient polyploidization, hybridization, dysploidy, and aneuploidy events (De Assis *et al.* 2013; Moraes *et al*. 2013; Cordeiro *et al*. 2022; Moraes et al. 2022; Nollet *et al*. 2022). However, ploidy and chromosome counts of both *E. anisatum*

and *E. marmoratum* are the same (2*n* = 40) and are consistent with previous diploid counts reported for the genus (2*n* = 40; Felix and Guerra 2010). Therefore, it appears that repetitive DNA dynamics is the primary cause of the approximately 2.3-fold genome size difference between these two species. Class I/Ty3-gypsy retrotransposons significantly contributed to the approximately 600 Mbp size difference. However, the largest difference in proportion is in the abundance of the *AniS1* satellite, with a difference of more than 11 percentage points.

The cluster corresponding to the *AniS1* satellite had a very good C index score (connected component index), whereas its P index (pair completeness index) score was relatively low, suggesting that *AniS1* is heavily dispersed throughout the *E. anisatum* genome. If this is the case, *AniS1* may be between the processes of dissemination and clustering (Ruiz-Ruano *et al*. 2016). Future fluorescence in situ hybridization (FISH) and third-generation sequencing studies will permit us to determine the distribution of *AniS1* across the *E. anisatum* genome.

Ty3-gypsy retrotransposon and satellite proportions have been identified as the primary factors contributing to genome size variation in other plants. In Pleurothallidinae, the proportion of Ty3-gypsy retrotransposons seems to explain genome size variation (Chumová *et al*. 2021). However, in other plant groups, high percentages of satellite DNA have been found in species with large genome sizes (i.e. He *et al*. 2015; Pellicer *et al*. 2021). Future work should investigate to what extent LTR transposons and satellite DNA have been responsible for shaping genome size variation in different lineages of *Epidendrum*, analyzing a greater portion of its taxic diversity in an evolutionary context.

CONCLUSIONS

Genomic profiling analysis of non-model species is an essential avenue for understanding the evolutionary history and genome dynamics of diverse lineages and is an essential step in high

coverage whole genome sequencing and assembly of reference genome. This is particularly relevant in plant lineages where even related species may exhibit dramatic differences in genome size. In this study, we present the first analysis of the repetitive DNA composition in two species of the orchid genus *Epidendrum*, examining heterozygosity rates, ploidy and repeat content. We observed a 2.3-fold difference in genome size with no evidence of polyploidy but revealing a predominance of Ty3-gypsy Ogre retrotransposons and a unique satellite DNA, jointly responsible for the difference in genome size between the species analyzed.

This study also offers insights on the comparison of genome size estimations derived from flow cytometry and k-mer analysis, noting their respective strengths, consistency, and limitations in the context of the genomic characteristics of *Epidendrum*, including the apparent absence of strict partial endoreplication, heterozygosity variation, and a high percentage of repetitive DNA. Our findings underscore the importance of tailoring bioinformatic methods to the particular characteristics of each taxonomic group and establish a baseline for further research focused on a better understanding of the size and composition of the genomes of *Epidendrum* and their possible role in the diversification of this genus.

Data Availability

All .histo files for genome size estimation, CLUSTER_TABLE.csv and SUPERCLUSTER_TABLE.csv files and Scripts are available at figshare: https://doi.org/10.6084/m9.figshare.c.8035945. Sequencing data analyzed for repeatome characterization is available at SRA (BioProject: PRJNA1330383).

Acknowledgments

We thank the Laboratorio de Biología Molecular, Instituto de Biología, Universidad Nacional Autónoma de México (UNAM) for access to DNA laboratory facilities; Eduardo A. Pérez García and the Orquidario Miguel Ángel Soto Arenas of the Facultad de Ciencias, UNAM, for providing material for cytometry analyses; and Cristian Cervantes, David Velázquez and Alfredo Wong from the Unidad de Sistemas y Tecnologías de la Información y Cómputo, Instituto de Biología, UNAM for support in using the HPC infrastructure. Access to this HPC cluster was made possible by the Instituto de Biología, UNAM, through Project Fronteras de la Ciencia 2016-01-1867, financed by Mexico's Consejo Nacional de Ciencia y Tecnología (CONACYT). Computational resources for the RepeatExplorer analysis were provided by the ELIXIR-CZ project (LM2023055), part of the international ELIXIR infrastructure.

Funding

This work was funded by the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica, Dirección General del Personal Académico, Universidad Nacional Autónoma de México [project IN221924, to GAS] and by the Consejo Nacional de Humanidades, Ciencias y Tecnologías [Fronteras de la Ciencia 2016, Project 1867, to SM]. Additional funding was provided by Instituto Chinoin, A.C.

Conflict of Interest

The authors declare no conflicts of interest.

References

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Britten, R. J. & Kohne D. E. (1968). Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science,* 161(3841), 529–540. https://doi.org/10.1126/science.161.3841.529

Brown, S. C., Bourge, M., Maunoury, N., Wong, M., Wolfe Bianchi, M., Lepers-Andrzejewski, S., Besse, P., Siljak-Yakovlev, S., Dron, M., & Satiat-Jeunemaître, B. (2017). DNA remodeling by strict partial endoreplication in orchids, an original process in the plant kingdom. *Genome Biology and Evolution*, *9*(4), 1051–1071. https://doi.org/10.1093/gbe/evx063

Chumová, Z., Záveská, E., Hloušková, P., Ponert, J., Schmidt, P. A., Čertner, M., Mandáková, T. & Trávníček, P. (2021). Repeat proliferation and partial endoreplication jointly shape the patterns of genome size evolution in orchids. *The Plant Journal: for cell and molecular biology*, 107(2), 511–524. https://doi.org/10.1111/tpj.15306

Chung, O., Kim, J., Bolser, D., Kim, H. M., Jun, J. H., Choi, J. P., Jang, H. D., Cho, Y. S., Bhak, J. & Kwak, M. (2022). A chromosome-scale genome assembly and annotation of the spring orchid (*Cymbidium goeringii*). *Molecular Ecology Resources*, 22(3), 1168–1177. https://doi.org/10.1111/1755-0998.13537

Cordeiro, J., Chase, M., Hágsater, E., Mendonça, E., Costa, L., Souza, G., Nollet, F. & Felix L. (2022) Chromosome number, heterochromatin, and genome size support recent polyploid origin of the *Epidendrum nocturnum* group and reveal a new species (Laeliinae, Orchidaceae). *Botany*, 100(5): 409–421. https://doi.org/10.1139/cjb-2021-0113

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M. & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2). https://doi.org/10.1093/gigascience/giab008

De Assis, F. N. M., Souza, B. C. Q., Medeiros-Neto, E., Pinheiro, F., Silva, A. E. B. & Felix, L. P. (2013). Karyology of the genus *Epidendrum* (Orchidaceae: Laeliinae) with emphasis on subgenus Amphiglottium and chromosome number variability in *Epidendrum secundum*: Karyology in *Epidendrum*. *Botanical Journal of the Linnean Society*, 172(3), 329–344. https://doi.org/10.1111/boj.12045

Dodsworth, S., Leitch, A. R. & Leitch, I. J. (2015). Genome size diversity in angiosperms and its influence on gene space. *Current Opinion in Genetics & Development*, 35: 73–78. https://doi.org/10.1016/j.gde.2015.10.006

Doležel, J., Greilhuber, J., Lucretti, S., Meister, A., Lysak, M., Nardi, L., y Obermayer, R. (1998). Plant Genome Size Estimation by Flow Cytometry: Inter-laboratory Comparison. *Annals of Botany*, 82, 17-26. https://doi.org/10.1006/anbo.1998.0730

Dominguez Del Angel, V., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Vinnere Pettersson, O., Amselem, J., Bouri, L., Bocs, S., Klopp, C., Gibrat, J.-F., Vlasova, A., Leskosek, B. L., Soler, L., Binzer-Panchal, M., & Lantz, H. (2018). Ten steps to get started in Genome Assembly and Annotation. F1000Research, 7, 148. https://doi.org/10.12688/f1000research.13598.1

Felix, L. P., & Guerra, M. (2010). Variation in chromosome number and the basic number of subfamily Epidendroideae (Orchidaceae). *Botanical Journal of the Linnean Society*, 163, 234-278. https://doi.org/10.1111/j.1095-8339.2010.01059.x

Giraud, D., Lima, O., Huteau, V., Coriton, O., Boutte, J., Kovarik, A., Leitch, A. R., Leitch, I. J., Aïnouche, M., & Salmon, A. (2021). Evolutionary dynamics of transposable

elements and satellite DNAs in polyploid *Spartina* species. *Plant Science*, 302, 110671. https://doi.org/10.1016/j.plantsci.2020.110671

Granados Mendoza, C., Jost, M., Hágsater, E., Magallón, S., Van Den Berg, C., Lemmon, E. M., Lemmon, A. R., Salazar, G. A., & Wanke, S. (2020). Target nuclear and off-target plastid hybrid enrichment data inform a range of evolutionary depths in the orchid genus *Epidendrum*. *Frontiers in Plant Science*, 10, 1761. https://doi.org/10.3389/fpls.2019.01761

Haberer, G., Young, S., Bharti, A. K., Gundlach, H., Raymond, C., Fuks, G., Butler, E., Wing, R. A., Rounsley, S., Birren, B., Nusbaum, C., Mayer, K. F. X., & Messing, J. (2005). Structure and Architecture of the Maize Genome. *Plant Physiology*, 139(4), 1612–1624. https://doi.org/10.1104/pp.105.068718

Hardie, D. C., Gregory, T. R. & Hebert, P.D. (2002). From pixels to picograms: a beginners' guide to genome quantification by Feulgen image analysis densitometry. *Journal of Histochemistry & Cytochemistry*, 50(6): 735–749. https://doi.org/10.1177/002215540205000601

He, Q., Cai, Z., Hu, T., Liu, H., Bao, C., Mao, W., & Jin, W. (2015). Repetitive sequence analysis and karyotyping reveals centromere-associated DNA sequences in radish (Raphanus sativus L.). *BMC Plant Biology*, 15(1), 105. https://doi.org/10.1186/s12870-015-0480-y

Hesse, U. (2023). K-Mer-based genome size estimation in theory and practice. In: Heitkam T, Garcia S. (eds.) *Plant Cytogenetics and Cytogenomics: Methods and Protocols*. Springer Nature: 79–113. https://doi.org/10.1007/978-1-0716-3226-0_4

Hloušková, P., Mandáková, T., Pouch, M., Trávníček, P. & Lysak, M. A. (2019). The large genome size variation in the Hesperis clade was shaped by the prevalent proliferation

of DNA repeats and rarer genome downsizing. *Annals of Botany*, 124(1): 103-120. https://doi.org/10.1093/aob/mcz036

Hozza, M., Vinař, T. & Brejová, B. (2015). How big is that genome? Estimating genome size and coverage from k-mer abundance spectra. In: Iliopoulos C, Puglisi S & Yilmaz E. (eds.) *String Processing and Information Retrieval: 22nd International Symposium*, SPIRE 2015, London, UK, September 1-4, 2015, Proceedings 22 (pp. 199-209). Springer International Publishing. https://doi.org/10.1007/978-3-319-23826-5_20

Hsu, C. C., Su, C. J., Jeng, M. F., Chen, W. H., y Chen, H. H. (2019). A HORT1 Retrotransposon Insertion in the PeMYB11 Promoter Causes Harlequin/Black Flowers in Phalaenopsis Orchids. *Plant Physiology*, 180(3), 1535–1548. https://doi.org/10.1104/pp.19.00205

Jenike, K.M., Campos-Domínguez, L., Boddé, M., Cerca, J., Hodson, C. N., Schatz, M. C. & Jaron, K.S. (2025). K-mer approaches for biodiversity genomics. *Genome Research* 35(2), 219–230. https://doi.org/10.1101/gr.279452.124

Karremans, A. P. (2021) With great biodiversity comes great responsibility: the underestimated diversity of *Epidendrum* (Orchidaceae). *Harvard Papers in Botany*, 26(2): 299–369. https://doi.org/10.3100/hpib.v26iss2.2021.n1

Kejnovsky, E., Hawkins, J.S., Feschotte, C. (2012). Plant Transposable Elements: Biology and Evolution. In: Wendel, J., Greilhuber, J., Dolezel, J., Leitch, I. (eds) *Plant Genome Diversity Volume 1*. Springer, Vienna. https://doi.org/10.1007/978-3-7091-1130-7_2

Lee, S. I. & Kim, N. S (2014). Transposable elements and genome size variations in plants. *Genomics & Informatics*, 12(3), 87-97. https://doi.org/10.5808/GI.2014.12.3.87

Leitch, I. J., Johnston, E., Pellicer, J., Hidalgo, O. & Bennett, M. D. (2019). Plant DNA C-values database (release 7.1, Apr 2019) https://cvalues.science.kew.org/

Leitch, I. J., Kahandawala, I., Suda, J., Hanson, L., Ingrouille, M. J., Chase, M. W. & Fay, M. F. (2009). Genome size diversity in orchids: Consequences and evolution. *Annals of Botany*, 104(3), 469–481. https://doi.org/10.1093/aob/mcp003

Lander, E. S. & Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, 2(3): 231–239. https://doi.org/10.1016/0888-7543(88)90007-9

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997v2*. https://doi.org/10.48550/arXiv.1303.3997

Li, S.-F., Su, T., Cheng, G.-Q., Wang, B.-X., Li, X., Deng, C.-L., & Gao, W.-J. (2017). Chromosome Evolution in Connection with Repetitive Sequences and Epigenetics in Plants. Genes, 8(10), 290. https://doi.org/10.3390/genes8100290

Liao, X., Zhu, W., Zhou, J., Li, H., Xu, X., Zhang, B. & Gao, X. (2023). Repetitive DNA sequence detection and its role in the human genome. *Communications Biology*, 6(1): 954. https://doi.org/10.1038/s42003-023-05322-y

Loureiro, J., Čertner, M., Lučanová, M., Sliwinska, E., Kolář, F., Doležel, J., Garcia, S., Castro, S. & Galbraith, D. W. (2023). The Use of Flow Cytometry for Estimating Genome Sizes and DNA Ploidy Levels in Plants. *Methods in Molecular Biology*, 2672, 25–64. https://doi.org/10.1007/978-1-0716-3226-0_2

Ma, J., Devos, K. M., & Bennetzen, J. L. (2004). Analyses of LTR-Retrotransposon Structures Reveal Recent and Rapid Genomic DNA Loss in Rice. *Genome Research*, 14(5), 860–869. https://doi.org/10.1101/gr.1466204

Macas, J., & Neumann, P. (2007). Ogre elements—A distinct group of plant Ty3/gypsy-like retrotransposons. *Gene*, 390(1–2), 108–116. https://doi.org/10.1016/j.gene.2006.08.007

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(11), e0143424. https://doi.org/10.1371/journal.pone.0143424

Marçais, G. & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6): 764-770. https://doi.org/10.1093/bioinformatics/btr011

Mehrotra, S. & Goyal, V. (2014). Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function. *Genomics, Proteomics & Bioinformatics*, 12(4), 164-171. https://doi.org/10.1016/j.gpb.2014.07.003

Moonlight, P. W., Baldaszti, L.,Cardoso, D., Elliott, A., Särkinen, T. & Knapp, S. (2024). Twenty years of big plant genera. *Proceedings of the Royal Society B*, 291: 20240702. https://doi.org/10.1098/rspb.2024.0702

Moraes, A. P., Chinaglia, M., Palma‐Silva, C., & Pinheiro, F. (2013). Interploidy hybridization in sympatric zones: The formation of *Epidendrum fulgens* × *E. puniceoluteum* hybrids (Epidendroideae, Orchidaceae). *Ecology and Evolution*, 3(11), 3824–3837. https://doi.org/10.1002/ece3.752

Moraes, A. P., Engel, T. B. J., Forni-Martins, E. R., De Barros, F., Felix, L. P. & Cabral, J.S. (2022). Are chromosome number and genome size associated with habit and environmental niche variables? Insights from the Neotropical orchids. *Annals of Botany*, 130(1): 11-25. https://doi.org/10.1093/aob/mcac021

Neumann, P., Novák, P., Hoštáková, N., & Macas, J. (2019). Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein

domains and provides a reference for element classification. *Mobile DNA*, 10(1), 1. https://doi.org/10.1186/s13100-018-0144-1

Nollet, F., Medeiros Neto, E., Cordeiro, J. M. P., Buril, M. T., Chase, M. W., & Felix, L. P. (2022). Chromosome numbers and heterochromatin variation in introgressed and non-introgressed populations of Epidendrum (Orchidaceae: Epidendroideae): interspecific transfers of heterochromatin lead to divergent variable karyotypes in the parental populations. *Botanical Journal of the Linnean Society*, 199(3), 694–705. https://doi.org/10.1093/botlinnean/boab101

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(12):e111. https://doi.org/10.1093/nar/gkx257

Novák, P., Neumann, P., & Macas, J. (2020). Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. *Nature Protocols*, 15(11), 3745–3776. https://doi.org/10.1038/s41596-020-0400-y

Palmer, M. J., Black IV, W. C. (1997). Determination of insect genome size and complexity by reassociation kinetics. *In* Crampton JM, Beard CB & Louis C. (eds.) *The Molecular Biology of Insect Disease Vectors: A Methods Manual*. Dordrecht: Springer Netherlands 172-194. https://doi.org/10.1007/978-94-009-1535-0_16

Pellicer, J., Fernández, P., Fay, M. F., Michálková, E., & Leitch, I. J. (2021). Genome Size Doubling Arises From the Differential Repetitive DNA Dynamics in the Genus Heloniopsis (Melanthiaceae). *Frontiers in Genetics*, 12, 726211. https://doi.org/10.3389/fgene.2021.726211

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

Pessoa, E. M., Cordeiro, J. M., Felix, L. P., Almeida, E. M., Costa, L., Nepomuceno, Á., Souza, G., Chase, M. W., Alves, M., & Van den Berg, C. (2021). Too many species: morphometrics, molecular phylogenetics and genome structure of a Brazilian species complex in *Epidendrum* (Laeliinae; Orchidaceae) reveal fewer species than previously thought. *Botanical Journal of the Linnean Society*, 195(2), 161-188. https://doi.org/10.1093/botlinnean/boaa071

Pflug, J. M., Holmes, V. R., Burrus, C., Johnston, J. S. & Maddison, D. R. (2020) Measuring genome sizes using read-depth, k-mers, and flow cytometry: methodological comparisons in beetles (Coleoptera). *G3: Genes, Genomes, Genetics*, 10(9): 3047-3060. https://doi.org/10.1534/g3.120.401028

Piegu, B., Guyot, R., Picault, N., Roulin, A., Saniyal, A., Kim, H., Collura, K., Brar, D. S., Jackson, S., Wing, R. A., & Panaud, O. (2006). Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. *Genome Research*, 16(10), 1262–1269. https://doi.org/10.1101/gr.5290206

Pinheiro, F. & Cozzolino, S. (2013) *Epidendrum* (Orchidaceae) as a model system for ecological and evolutionary studies in the Neotropics. *Taxon*, 62: 77-88. https://doi.org/10.1002/tax.621007

R Core Team (2024) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.

Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. (2020) GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* 11(1): 1432. https://doi.org/10.1038/s41467-020-14998-3

Ruiz-Ruano, F. J., López-León, M. D., Cabrero, J., & Camacho, J. P. M. (2016). High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Scientific Reports*, 6(1), 28333. https://doi.org/10.1038/srep28333

Shu, Z., Row, S., & Deng, W. M. (2018). Endoreplication: The Good, the Bad, and the Ugly. *Trends in Cell Biology*, 28(6), 465–474. https://doi.org/10.1016/j.tcb.2018.02.006

Sun, H., Ding, J., Piednoël, M. & Schneeberger, K. (2018). findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics* 34(4): 550-557. https://doi.org/10.1093/bioinformatics/btx637

Trávníček, P., Ponert, J., Urfus, T., Jersáková, J., Vrána, J., Hřibová, E., Doležel, J., & Suda, J. (2015). Challenges of flow-cytometric estimation of nuclear genome size in orchids, a plant group with both whole-genome and progressively partial endoreplication. *Cytometry Part A*, 87(10), 958–966. https://doi.org/10.1002/cyto.a.22681

Vicient, C. M., Jääskeläinen, M. J., Kalendar, R., & Schulman, A. H. (2001). Active retrotransposons are a common feature of grass genomes. *Plant Physiology*, 125(3), 1283–1292. https://doi.org/10.1104/pp.125.3.1283

Vitte, C., Panaud, O., & Quesneville, H. (2007). LTR retrotransposons in rice (*Oryza sativa*, L.): Recent burst amplifications followed by rapid DNA loss. *BMC Genomics*, 8(1), 218. https://doi.org/10.1186/1471-2164-8-218

Vurture, G.W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., Schatz, M. C. (2017). GenomeScope: fast reference-free genome profiling from short

reads. *Bioinformatics*, 33(14), 2202–2204. https://doi.org/10.1093/bioinformatics/btx153

Wang, D., Zheng, Z., Li, Y., Hu, H., Wang, Z., Du, X., Zhang, S., Zhu, M., Dong, L., Ren, G., & Yang, Y. (2021). Which factors contribute most to genome size variation within angiosperms? Ecology and Evolution, 11(6), 2660–2668. https://doi.org/10.1002/ece3.7222

Weiß, C. L., Pais, M., Cano, L. M., Kamoun, S., & Burbano, H. A. (2018). nQuire: a statistical framework for ploidy estimation using next generation sequencing. *BMC Bioinformatics*, 19(1), 122. https://doi.org/10.1186/s12859-018-2128-z

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Figures

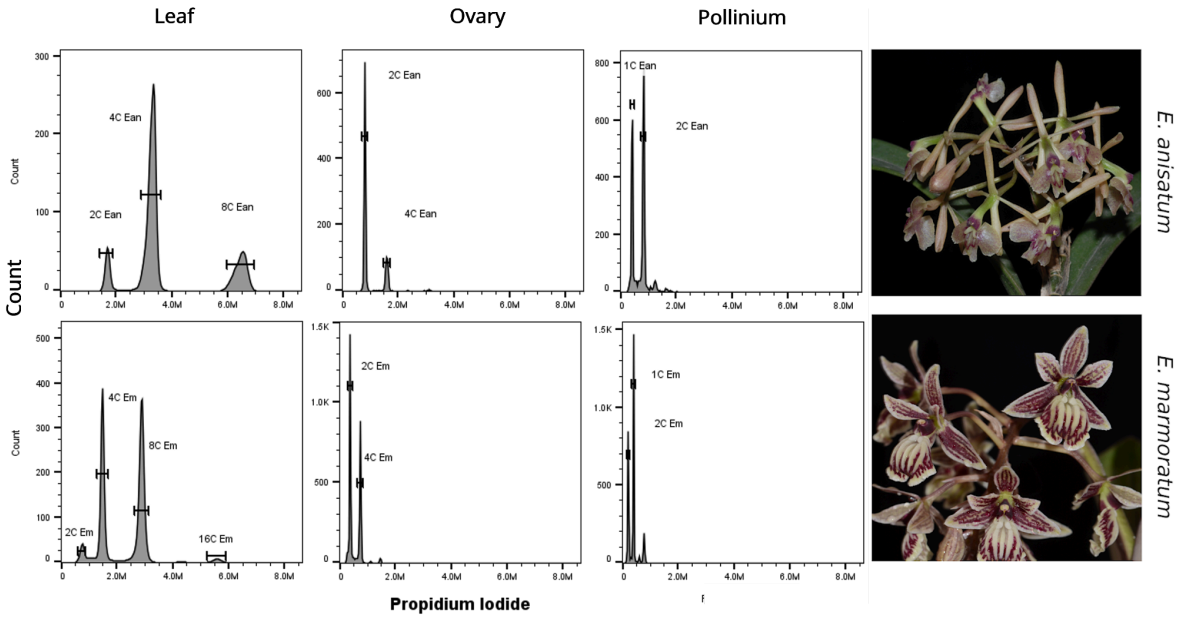

Fig 1. DNA content histograms for *Epidendrum anisatum* (voucher: *Salazar et al. 7370*; upper right) and *E. marmoratum* (voucher: *Salazar 10329*; lower right) from three different tissues: leaf (left), ovary (middle), and pollinium (right). Abbreviations: Ean, *E. anisatum*; Em, *E. marmoratum*.

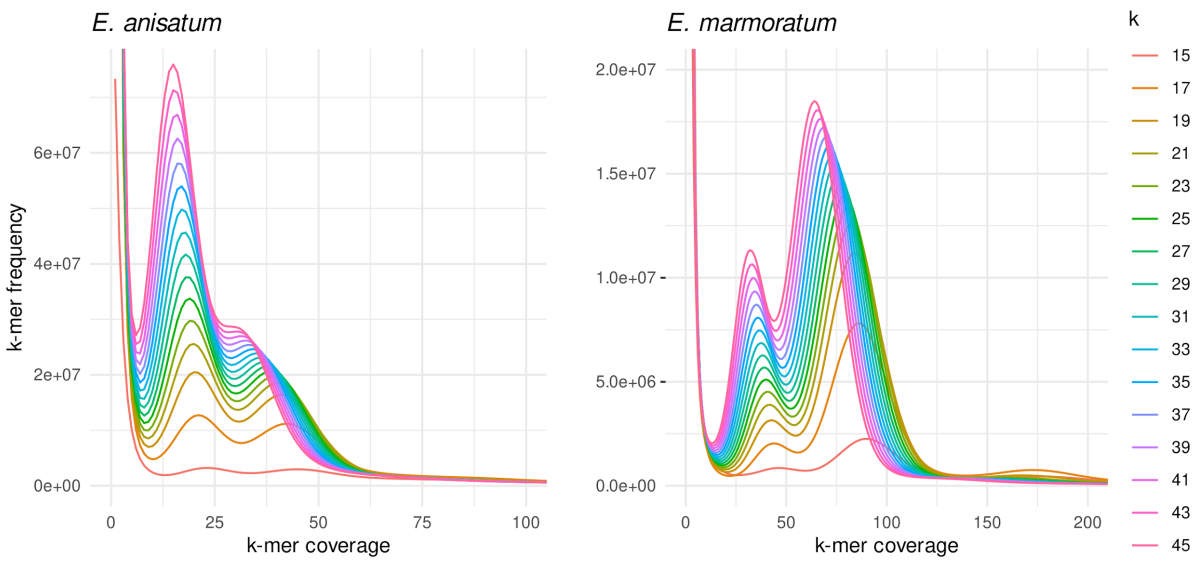

Fig 2. K-mer analysis histograms for *Epidendrum anisatum* (left) and *E. marmoratum* (right) for 16 values of k.

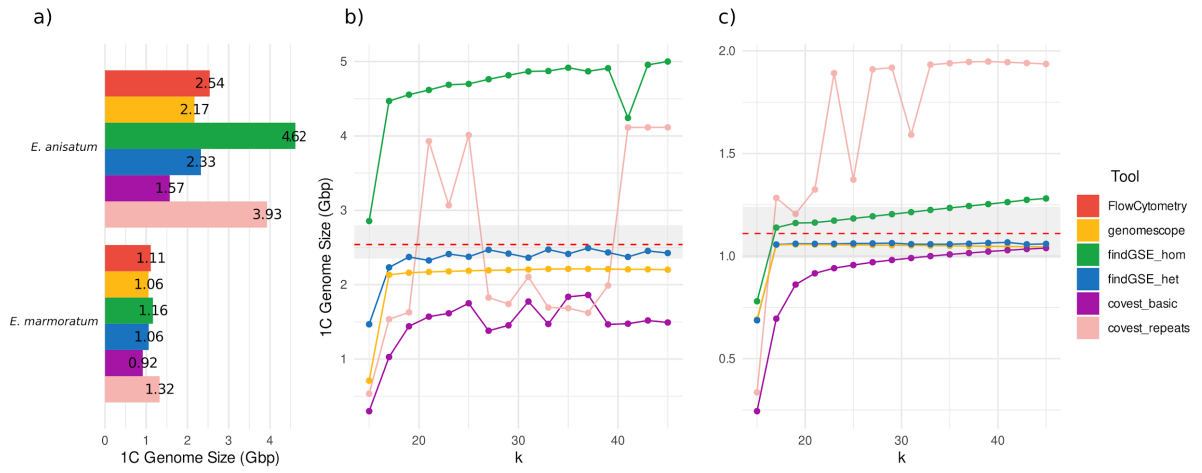

Fig 3. Genome size estimates for *Epidendrum anisatum* and *E. marmoratum* with flow cytometry and five bioinformatics tools. a) Comparison for k = 21, b) genome size estimates for *E. anisatum* with multiple values of k; c) genome size estimates for *E. marmoratum* with multiple values of k. The grey rectangle represents the lowest and highest flow cytometry estimates, and the red dotted line represents the average flow cytometry value.

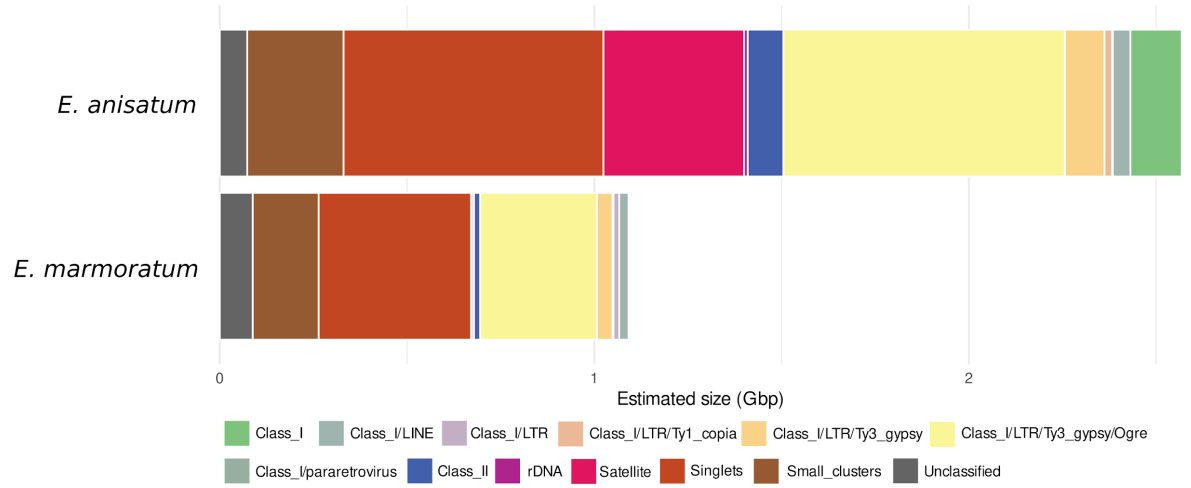

Fig. 4. Abundance of repetitive DNA for the two species of *Epidendrum* analyzed after automatic annotation and manual review in RepeatExplorer2.

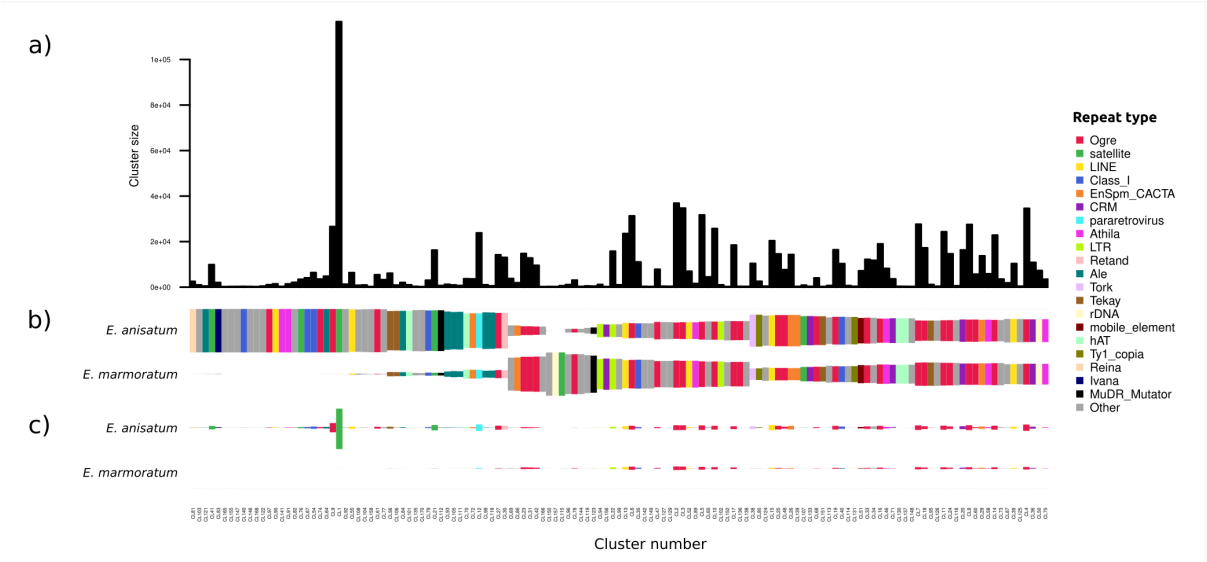

Fig 5. Results of the comparative protocol in RepeatExplorer2. a) Cluster size plot for all clusters, b) proportion comparison per cluster for *Epidendrum anisatum* and *E. marmoratum*, the rectangles represent what proportion of the cluster belongs to each species, c) abundance comparison per cluster for *E. anisatum* and *E. marmoratum*. In this case, the rectangles represent the absolute abundance of each cluster for each species.

Supplementary Data Fig. S1. K-mer analysis histograms for *Epidendrum anisatum* (upper row) and *E. marmoratum* (lower row) with 16 different values of k for five different depth of coverage sampling values (5×, 10×, 20×, 30×, and 40×).

Supplementary Data Fig. S2. Genome size estimates for *Epidendrum anisatum* and *E. marmoratum* using histograms generated from a k-mer counting with a maximum k-mer coverage value of 10,000 using 16 different values of k.

Supplementary Data Fig. S3. Genome size estimates for *Epidendrum anisatum* and *E. marmoratum* with flow cytometry and five bioinformatics tools using 16 different values of k

and 5 different depth of coverage sampling values (5×, 10×, 20×, 30×, and 40×). An estimate of zero means that the program did not produce a result for that combination of values.

Tables

Table 1. nQuire model comparisons

Species	Free model log-likelihood	Δ Diploid log-likelihood	Δ Triploid log-likelihood	Δ Tetraploid log-likelihood
E. anisatum	9443.888813	**576.308038**	3668.262772	4422.734976
E. marmoratum	3556.02657	**644.05066**	1310.933689	1422.859328

Table 2. Detailed repeat composition breakdown and comparison by genome proportion.

Sequence type	Transposon classification	*E. anisatum* (%)	*E. marmoratum* (%)
Singlets		***27***	***37***
Repeats		***73***	***63***
Unclassified repeats		**2.85**	**7.99**
Transposons		**45.1**	**37.93**
	Class I	**41.4**	**36.43**
	Class I Unclassified	*5.33*	*0.43*
	LINE	*1.83*	*2.28*
	LTR	*34.24*	*33.72*
	LTR Unclassified	*0.07*	*1.37*
	Ty1-copia	*0.8*	*0.25*
	Ty3-gypsy	*33.37*	*32.1*
	Ty3-gypsy Unclassified	0.58	0
	chromovirus/CRM	1.3	2.13
	chromovirus/Reina	0.24	0.001
	chromovirus/Tekay	0.13	0.21
	non-chromovirus/Athila	0.73	1.11
	non-chromovirus/Ogre	29.24	28.34
	non-chromovirus/Retand	1.15	0.31
	Class_II	**3.7**	**1.5**
	Subclass_1	**3.7**	**1.5**
	TIR Unclassified	0.31	0
	EnSpm_CACTA	2.74	0.58
	hAT	0.65	0.8
	MuDR_Mutator	0	0.11
rDNA		**0.39**	**0.51**
Satellite		**14.6**	**0.19**
Small clusters		**10**	**16**

*Only showing repeat types with at least 0.1% genome proportion in at least one of the two species

** After both automatic annotation and manual review