universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Evolution of genome sizes and repetitive DNA fractions in selected species in the genus Tillandsia

verfasst von | submitted by

Neil Andrew McNair MPhys PhD

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 910
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Computational Science
Betreut von | Supervisor:	Univ.-Prof. Mag. Dr. Hanna Schneeweiss Privatdoz.

Table of Contents

List of Tables

List of Figures

1 Introduction

Repetitive DNA sequences in eukaryotic genomes are ubiquitous. In all genomes examined to date, several major DNA sequence types are found to occur multiple times [Luo et al., 2022]. The proportion of a genome that is composed of repeats can be very high, from 54% in humans [Nurk et al., 2022] to above 80% in some plants. *Zea mays* for example, has a repeat abundance of between 77% and 81% depending on the variety [Haberer et al., 2020], *Hordeum vulgare* has 84% [Mayer et al., 2012]. Large genome sizes are associated with high repeat abundances, as shown in *Pinus taeda* (20.6 Gb, 82%) [Wegrzyn et al., 2014], and *Axolotl mexicanum* (32.4 Gb, 65.6% repeats) [Nowoshilow et al., 2018]. While some repetitive DNA types are coding, such as the tandem repeats of rDNA, most, such as satellite DNAs, are non-coding, and some, such as transposable elements, code only for proteins required for their own propagation in the genome [Wicker et al., 2007].

Repetitive DNA has long been considered to represent "junk DNA" or parasitic DNA, raising the question of why such DNA persists and is maintained in the genome. The modern paradigm is to view this material as "genomic dark matter", being until recently excluded from genomic assemblies, which formerly relied upon short read sequence data, but nevertheless being present in large quantities, having an important influence on the functioning of the organism (see e.g. [Shapiro & Sternberg, 2005]) - just as astronomical dark matter is invisible to traditional optical telescopes, but which is now beginning to be better understood thanks to newer observational technologies.

1.1 The genus *Tillandsia*

Tillandsia is a diverse genus of bromeliads (Bromeliaceae) within the subfamily Tillandsioideae, native to a large part of the Americas. The distribution range of *Tillandsia* extends from the southern USA, with the well-known Spanish moss, *Tillandsia usneoides*, to northern Argentina. Most *Tillandsia* species are epiphytes,

while some are lithophytes or aerophytes, living attached either to other plants or to rocks, rather than rooted in the soil. Lacking true roots, they obtain water from rain or from condensation of water vapour. The genus is characterized by diverse morphologies, habitats, and types of photosynthesis [Barfuss et al., 2016].

There are more than 700 *Tillandsia* species, with a great diversity of growth habits, and several recurrently evolving phenotypic traits. Some species use C3 photosynthesis, like most plants, and some use CAM (Crassulacean Acid Metabolism) photosynthesis which is an adaptation to hot and dry daytime conditions. Some species have smooth leaves and rainwater tanks formed by the arrangement of their leaves as an adaption to store water, while others have trichomes on their leaves to collect moisture from condensation. These different adaptations allow *Tillandsia* species to occupy a variety of ecological niches, some species living in wet conditions and some in dry, some on trees (as epiphytes) and some on rocks (as lithophytes). The transition between C3 and CAM occurred at multiple points in the *Tillandsia* phylogeny [de la Harpe et al., 2020].

The phylogenetic relationships of *Tillandsia* have recently been elucidated [Barfuss et al., 2016], and the genus is divided into seven subgenera, including *Tillandsia subg. Tillandsia*. The great diversity within the genus has arisen within the past few million years, and is attributed to an adaptive radiation - a rapid emergence and diversification of many species [Benzing, 2000]. The *Tillandsia* radiation has been dated to around 14 million years ago, with its point of origin inferred to be in South America, in the Andes [Givnish et al., 2011]. Newly evolved *Tillandsia* species rapidly adapted to fill diverse ecological niches and extended their geographical ranges. Diversification within South America was followed by an extension of the geographic range of the genus to Central America, Mexico, the Caribbean islands, and parts of the southern USA.

Only 35 *Tillandsia* species have had their genome sizes measured using flow cytometry [Zonneveld et al., 2005; Favoreto et al., 2012; Gitai et al., 2014; Müller et al., 2019]. The measured genome sizes range from 1C = 391 to 1,633 Mb. Chromosome numbers in *Tillandsia* have been established for 94 species [Brown & Gilmartin 1989;

review in Gitai et al., 2014]. Most *Tillandsia* species have $2n = 50$, a character shared with the bromeliad "model species" *Ananas comosus*. An exception known to date is *Tillandsia leiboldiana*, with $2n = 38$ [Brown & Gilmartin, 1989; Groot Crego et al., 2024]. The wide range of genome sizes in *Tillandsia*, accompanying mostly stable chromosome numbers, suggests that the larger genomes may have higher abundances of repeats. However, this has not been previously analysed, with the exception of repeat abundance estimates for two species, *T. fasciculata* and *T. leiboldiana*, obtained from whole genome assemblies [Groot Crego et al., 2024].

Two major processes contribute to genome size variation between related species - polyploidy and changes of copy numbers of various types of repetitive DNAs, including dispersed and tandem repeats. The most common chromosome number found in *Tillandsia* to date, $2n = 50$, is probably of (paleo)polyploid origin, suggesting the adaptive radiation of the genus took place sometime after a shared polyploidy event. More recent polyploidy has only been reported in the subgenus *Diaphoranthema* [Till 1992], with $2n = 4x = 100$. Therefore, the changes in genome size in Tillandsia species accompanying the adaptive radiation are most likely due to changes in the abundance of repetitive DNAs. This makes this genus an excellent system in which to study the dynamics of repeats and their contribution to genome size variation.

1.1.1 Species studied in this project

21 species were chosen for this project (16 species of *Tillandsia*, and 5 representing other genera in Tillandsioideae). They were selected to represent a range of ecotypes, photosynthetic types, and geographical areas – including both CAM and C3 species (11 and 10, respectively). Some species possess tanks (nine species) and some have succulent leaves and trichomes (eight species). Several species are endemic to a single country, and others are widespread. Only six of the analyzed species had known chromosome numbers. Three of these six taxa were reported to

have $2n = 50$, one was reported as $2n = 48$ or 50, one as $2n = 38$, and the last as $2n = 22$ (Table 1).

The phylogeny of Tillandsioideae is described in several previous papers [Barfuss et al., 2005; Barfuss et al., 2016]. Among the 16 *Tillandsia* species studied here, five (*T. fasciculata*, *T. ionantha*, *T. juncea*, *T. leiboldiana*, and *T. propagulifera*) are in *Tillandsia* subg. *Tillandsia*. Four of these (*T. fasciculata*, *T. ionantha*, *T. juncea*, and *T. leiboldiana*) represent clade K, the "Mexican clade", which is of especial interest as genomic assemblies have recently been published for two of these species (*T. fasciculata* and *T. leiboldiana*) [Groot Crego et al., 2024]. The choice of these particular species, from the ~700 *Tillandsia* species, was also guided by the availability of pre-existing short read sequence data [de la Harpe et al., 2020].

Table 1. Species studied in this project.

Species	Type of photo-synthesis	Morphology	Geographical range	2*n*	Genome size (Mb/1C)
T. australis *	C3	tank	Bolivia, Argentina		
T. aff. australis	C3	tank	Bolivia, Argentina		
T. complanata	C3	tank	Bolivia, Brazil to Costa Rica, Greater Antilles	22 (a)	
T. curvispica	CAM	succulence & trichomes	Peru		
T. demissa	CAM	tank	Ecuador		
T. aff. divaricata	CAM	succulence & trichomes	Ecuador	50 (a)	
T. fasciculata *	CAM	succulence & trichomes	Mexico, Central America, West Indies, northern South America	50 (a); 50 (d)	768 (c); 838 (d)
T. floribunda	CAM	succulence & trichomes	Ecuador, Peru		
T. ionantha *	CAM	succulence & trichomes	Mexico, Central America, Brazil		699 (c)

Species	Type of photo-synthesis	Morphology	Geographical range	2*n*	Genome size (Mb/1C)
T. juncea	CAM	succulence & trichomes	Bolivia to Mexico, Greater Antilles	50 (a)	1,139 (c)
T. latifolia	CAM	succulence & trichomes	Peru	50 (a); 48 (b)	509 (b)
T. leiboldiana *	C3	tank	Mexico & Central America	38 (a); 38 (d)	1,198 (d)
T. propagulifera *	CAM	succulence & trichomes	Peru		
T. rauhii	CAM	trichomes & tank	Peru, Central America		846 (c)
T. somnians	C3	tank	Ecuador, Peru		
T. sphaerocephala	CAM	tank	Peru, Bolivia, Argentina		
Barfussia laxissima	C3	tank	Bolivia		
Lemeltonia dodsonii	C3	no succulence, no tank	Ecuador		
Pseudalcantarea macropetala	C3	tank	Mexico		
Racinea spiculosa	C3	no succulence, no tank	from Brazil & Bolivia to Mexico		
Wallisia lindeniana	C3	no succulence, no tank	Ecuador		

The five species indicated with * were used for analyses of repetitive DNA fractions. Sources for chromosome numbers and genome sizes: (a) = Brown & Gilmartin, 1989; (b) = Gitai et al., 2014; (c) = Müller et al., 2019; (d) = Groot Crego et al., 2024.

In the second part of this project, the repetitive content of the genomes of five of the species is investigated in detail. Three of the five species represent the Mexican clade (*T. fasciculata*, *T. ionantha*, and *T. leiboldiana*), and are found in Mexico and Central America, while the other two taxa are found only in South America (*T. propagulifera* and *T. australis*). *T. propagulifera* is in subgenus Tillandsia, but outside the Mexican clade (clade K); *T. australis* is more distantly related and placed in the *T. australis* complex. Three of the five analysed species (*T. fasciculata*, *T. ionantha*, and *T.*

propagulifera) use CAM photosynthesis, and have succulent leaves with trichomes. In contrast, *T. leiboldiana* and *T. australis* use C3 photosynthesis and have tanks for collecting rainwater. Phylogenetically, the three CAM species are no more closely related to each other than they are to the others; neither are the two C3 species.

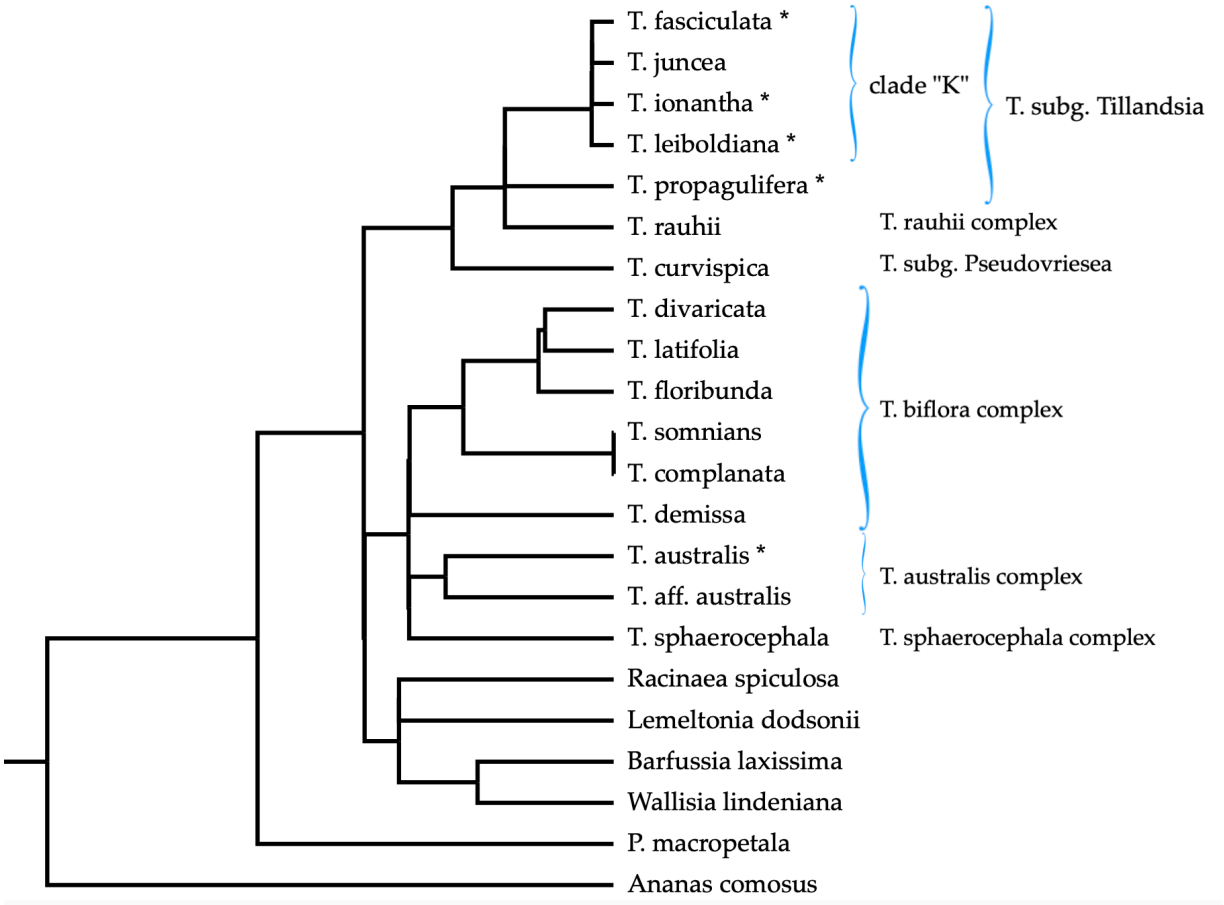


Figure 1. Phylogeny of the 21 analyzed species.
Ananas comosus is included as the outgroup. The clades within *Tillandsia* are labelled, and the five species whose repeatomes were investigated are marked with *. Adapted from de la Harpe et al., [2020].

1.2 Repetitive DNA

Two major types of repetitive DNA can be distinguished - tandem repeats and dispersed repeats. Tandem repeats are composed of DNA sequence motifs, or 'monomers', repeated numerous times in a head-to-tail orientation in one or more loci in the genome. Tandem repeats include both non-coding satellite DNAs and the only coding repetitive DNA sequence types in the genome, 5S and 35S rRNA genes

[Goffová & Fajkus, 2021]. Dispersed repeats, in contrast, are monomers scattered throughout the genome [Bennetzen & Wang, 2014]. Dispersed repeats include several classes and families of transposable elements, abbreviated as TEs. They are named after their ability to transpose in the genome, that is to move to a new genomic location. TEs were first discovered in *Zea mays* in the 1940s [McClintock, 1950] and they are found in all eukaryotic genomes analysed to date. Transposable elements make use of the host cell's genetic machinery to be transcribed and translated, and then use their own enzymes to transpose. Most genomes contain unique combinations of various TE types, some of them in thousands of copies [Quesneville, 2020]. Each transposition event may result in an increase in the size of the host genome, and it has been shown in many genomes that these processes are ongoing, for example in *Nicotiana* or *Hordeum* [Grandbastien et al., 1989; Gao & Fox-Fogle, 2023]. However, most TE copies are deactivated in subsequent generations, silenced by their host genome, and many eventually become non-autonomous, having lost the ability to transpose.

1.2.1 Classification of repeats

The classification of dispersed transposable elements (TEs) into classes, subclasses, and families depends upon the mode of their transposition, the structure of their flanking regions, and the types and order of the coding domains they carry. There are two classes of transposable elements - class I (retrotransposons) and class II (DNA transposons with terminal inverted repeats (TIRs) and *Helitrons*) [Wicker et al., 2007]; this classification has recently been extended, but these groups are major units used in classification by RepeatExplorer and are most abundant in plant genomes and will therefore be described here. The two classes are distinguished based on the mechanism they use to transpose. Each class is further subdivided into subclasses, orders, and superfamilies, based on the internal structure of the elements, i.e. the types of genes they carry, and on the form of the terminal regions.

Class I dispersed repeats, retrotransposons, use an RNA intermediate to transpose themselves. One of the coding regions that they carry codes for a reverse transcriptase. Once the original copy of the transposon (made of DNA) has been transcribed into mRNA and translated, the reverse transcriptase creates new cDNA copies of the element based on the mRNAs either directly at the integration site (non-LTR-elements) or in the cytoplasm followed by insertion of new copies into genomic locations in the nucleus by another enzyme, an integrase which is also encoded by the element itself (Long Terminal Repeat (LTR)-retrotransposons). This process is referred to as "copy-and-paste". The new copy is flanked by target site duplication (TSD) - a few bases from the target site now exist in two copies, on either side of the new element.

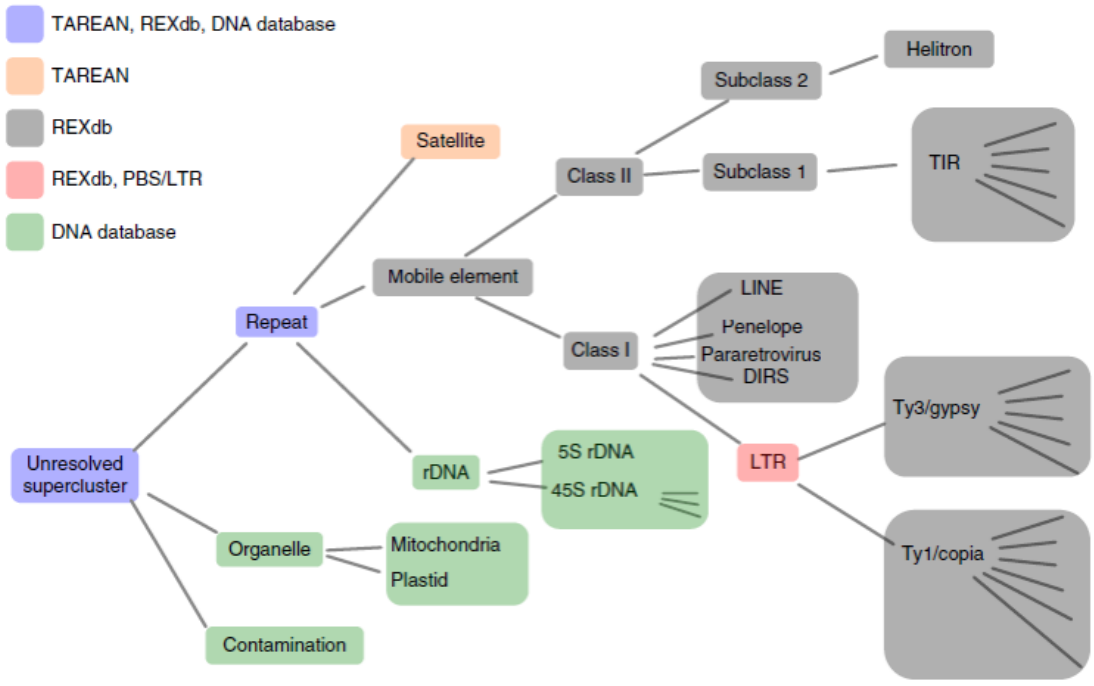


Figure 2. Basic classification of repeats in eukaryotic genomes.
This classification is used for annotation of the clustering results by RepeatExplorer (Novák et al., 2020).

The more common of the two subclasses are the LTR-retrotransposons, in which the coding domains of the TE are flanked by identical LTRs, in the same orientation, ranging in length from a few hundred up to a few thousand bases. Two major superfamilies of LTR-retrotransposons are Ty3-*gypsy* and Ty1-*copia*, which are the most common types of repeats in most land plants [Neumann et al., 2019]. The

second subclass of retrotransposons comprises LINEs and SINEs - long and short interspersed nuclear elements [Elbarbary et al., 2016]. These are not flanked by long terminal repeats (non-LTR). LINEs code for both a reverse transcriptase and an endonuclease, and are capable of transposing themselves; SINEs, however, lack these coding domains, and are 'non-autonomous', relying on the presence of related LINEs' machinery to be transposed.

In contrast to retrotransposons, class II transposable elements do not use an RNA intermediate to transpose. Class II consists of up to four subclasses, two of which are well understood and relevant for plants: subclass I, DNA transposons (TIRs), use a “cut and paste” mechanism where the element is cut out from its location in the genome and moved to a new location. These elements contain a coding domain that produces a transposase, an enzyme which excises the transposon from the genome, and then inserts it in a different location. The transposase is flanked by two Terminal Inverted Repeats (TIRs). Most of the transposition events do not result in increase of copy numbers of the DNA TEs. Occasionally, however, due to the gap repair mechanism the old copy of the TE in its original position may be retained. Subclass II TEs of class II use a "peel and paste" transposition mechanism that does not involve double strand cleavage [Wicker et al., 2007]. They include most notably *Helitrons*, which transpose using a 'rolling circle' mechanism, with a single strand DNA intermediate [Thomas & Pritham, 2015]. *Helitrons* contain a coding domain that produces Rep/Helicase, a protein that acts as both an endonuclease and as a helicase.

In addition to dispersed transposable elements the tandem repeats also significantly contribute to the host genome size. These mainly comprise ribosomal DNAs (rDNAs) and satellite DNAs. Ribosomal DNAs (housekeeping genes) code for the ribosomal RNAs that are the main building blocks of ribosomes. The nuclear genome of a eukaryotic cell contains many copies of these genes, so that many ribosomes can be quickly assembled when required. Although the genes themselves are highly conserved, the intergenic spacers (NTS for 5S rDNA; ITS1, ITS2, ETS, and IGS for 35S rDNA) are very variable and thus often used as phylogenetic markers

[Chen et al., 2016; Hu et al., 2019]. There are two types of rRNA genes: 5S rRNA and 35S rRNA genes. The latter codes for 18S, 5.8S, and 25S rRNAs [Rossello et al., 2022]. All eukaryotic genomes contain at least one locus for each of the two types of rDNA, but many species contain multiple loci of either of the rRNA genes [Goffová & Fajkus, 2021]. Each locus contains hundreds to tens of thousands of rRNA gene copies.

Satellite DNAs represent tandem repeats, but they are non-coding. Satellite DNAs are often species- or genus-specific and are highly variable in monomer sequence and in the number and position of the loci. They usually originate from existing low copy number repeated motifs in the genome and can change copy number and number of loci via unequal recombination and extrachromosomal circular DNAs. Satellite DNA monomers range in length from 2 or 3 bp up to thousands of bp and can mutate and homogenize within the arrays very rapidly. Although often found in and around centromeres or telomeres, satellite DNAs also occur in interstitial loci [Šatović-Vukšić & Plohl, 2023].

1.3 Genome size evolution and repetitive DNA

Among higher plant species, genome size varies by more than 2000-fold, ranging from a bacterial-size 61 Mb in *Genlisea tuberosa*, to 148,851 Mb in *Paris japoni*ca [Pellicer et al., 2010, Fleischmann et al., 2014]. This wide range gave rise to the so-called C-value paradox. The paradox was partially explained in the same decade it was first formulated, the 1970s, with the discovery that only a small part of an organism's genome is actually composed of genes required for the development and function of the organism itself. Thus, although the number of genes is rather similar for all species within each large phylogenetic group (e.g., among higher plants), the total genome size can vary widely depending on the amount of repetitive DNA [Bennetzen et al., 2005]. Given the material and energy required to replicate a large quantity of DNA at every cell division, there is a selection pressure towards removing unnecessary genetic material. For genomes with a high abundance of

repetitive DNAs, it has therefore been hypothesized that these repetitive DNAs have an adaptive role in their host genomes.

Evolutionary changes in the abundances of transposable elements and satellite DNAs contribute significantly to changes - both increases and decreases - of the genome size. Activation of retrotransposons can cause rapid increases of the genome size. These bursts of transposition may be followed by a longer period of genome size stability or genome size reduction. The processes responsible for the removal of repeats from a genome, are much less efficient than the amplification of transposable elements or tandem repeats and the consequent genome size increase. The processes of polyploidization and subsequent diploidization also bring about changes in genome size, but are accompanied by changes of chromosome number.

A comparative analysis of the abundances of different classes of repeats among a group of related species allows for the inference of the dynamics of genome size changes and of the mechanisms responsible. It has been suggested that a high level of TE activity leading to both increases and decreases in TE abundances may be associated with, and may indeed be one of the drivers of, an adaptive radiation, such as is seen in *Tillandsia* [Benzing, 2000] and in *Mikania* [Su et al., 2021], because the increase in genetic diversity allows the species to evolve faster and to take advantage of new ecological niches.

Reactivation of silenced TEs has been shown to occur as a response to various abiotic and biotic stress conditions [e.g. Sun et al., 2020]. Therefore TEs can play a role in adaption to stress, proliferating in the genome and increasing genetic diversity, which provides raw material for evolution and may facilitate the rapid evolution of species into new niches - an adaptive radiation. It has been hypothesized that the uplift of the Andes mountain chain may have been both the cause of environmental stress, and also the source of new niches to be occupied and adapted into by radiating *Tillandia* species [Luebert & Weigend, 2014].

1.4 Repeatome analysis in plants

The analysis of the repetitive part of the genome, the repeatome, is not straightforward due to the repetitiveness of the elements themselves. Following the introduction of next generation sequencing (NGS) technologies, new bioinformatic approaches were developed to identify, assemble, and annotate genomic repeats. Two approaches are particularly effective - the library approach and *de novo* identification by comparison between multiple reads in a sample. The library approach uses a curated library of repeating element sequences, often optimized for a particular species, as a source of comparison to identify elements in the sample being studied.

The second approach is used by RepeatExplorer [Novák et al., 2013 (version 1); Novák et al., 2020 (version 2, used here)]. This is a pipeline devoted to the analysis of the repeatome of any species using a genome skimming strategy and short read DNA data, without the need for a genome assembly, or any other pre-existing genomic information, other than an estimate of the genome size. The pipeline does not rely on the affinity of the study species to any other taxon. RepeatExplorer calculates the sequence similarity between all the sample reads, in order to group the reads together into clusters, each of which represents a major repeat type [Novák et al., 2010]. RepeatExplorer also annotates the repeats, using a curated library, REXdb [Neumann et al., 2019], of coding regions from many repeat families.

1.5 Estimation of genome sizes

The most accurate method for estimating the size of a genome is flow cytometry [Baranyi et al., 1996; Temsch, 2003]. This method, however, requires fresh plant material and is therefore not always feasible. Therefore, alternative bioinformatic methods have been developed, which make use of the wide availability of genomic sequence data. Two of these methods are the exon alignment method and the k-mer method.

1.5.1 The exon alignment method

This method determines the depth of coverage of sequence data from the studied species, by aligning that sequence data to a genome assembly of a related reference species [Pflug et al., 2020]. The assembly must be annotated because alignment with exons in the reference species is used to identify exonic sequences in the analysed species. Exons - the coding parts of genes - are usually highly conserved. Therefore, it is assumed that the exonic parts of two related genomes will be very similar, both in terms of their DNA sequences, and also their total length. This will be true even if the two genomes differ in their repeat content. Alignment of the short-read DNA sequence data between each of the study species and the reference species assembly, allows the estimation of the proportion of exonic regions in each dataset. Under the assumption that each exome in related species has the same size, the depth of coverage of each set of data can be estimated. In other words, dividing the total length of all the reads in the study species that align to exons in the known genome assembly, by the total length of the genome assembly exome, results in an estimate of the depth of coverage of the dataset of the target species. The estimated coverage, in turn, allows the calculation of the size of each genome. The advantage of this method is that it can be used for any species that has a related species which has an assembled and annotated genome.

1.5.2 The *k*-mer method

The second approach used to estimate genome size is the *k*-mer method [Li & Waterman, 2003; Hesse, 2023]. It uses the shape of the *k*-mer frequency spectrum to infer information about a genome. Like the exon alignment method, the *k*-mer method can be used for a genome where no assembly exists, however it does not rely on having an annotated genome assembly of a related species. A *k*-mer is a short DNA sequence motif, of length *k*. As *k*-mer length increases, the number of distinct

possible *k*-mers increases geometrically. In a completely random genetic sequence, all possible *k*-mers, for a given value of *k*, will be equally likely to occur. However, in real genomes some *k*-mers occur more frequently than others, providing information about the composition of the genome: its size, its heterozygosity, its level of ploidy, and the abundance of repeats.

Useful information about repeats in the genome is summarized in the *k*-mer frequency spectrum. For a chosen value of *k*, the number of occurrences of each distinct *k*-mer is counted. The next step is find how many *k*-mers occur once only, how many occur twice, how many occur three times, and so on. This set of numbers is the *k*-mer frequency spectrum. It is visualized as a histogram, with each column showing the abundance of *k*-mers of the corresponding multiplicity. The *k*-mer frequency spectrum might have only one peak, or several. The positions and relative heights of the other peaks provide information about depth of coverage, heterozygosity, and ploidy. The position of the main peak indicates the depth of coverage, the number of copies of the genome that are present in the sequenced dataset. The positions and relative heights of the second and further peaks provide information about heterozygosity and ploidy level, while the shape of the long tail to the right can be used to estimate the abundance of repeats in the genome.

To infer the genome size, the coverage depth of the dataset per base is needed. It is given by the following formula:

$$C_G = \frac{L_R \times C_k}{(L_R - k + 1)}$$

where C_G denotes the depth of coverage of the dataset, C_k is the *k*-mer depth of coverage, L_R is the length of each read, and k is the chosen *k*-mer length [Hesse, 2023]. When using this method, the choice of *k* is important. Firstly, *k* needs to be odd, to avoid double counting of palindromes. Secondly, the shape of the *k*-mer spectrum varies with the choice of *k*, which should be chosen so that the peaks are clearly distinguishable.

1.6 Aims

In this work two questions concerning genome size and repeat abundance in the genus *Tillandsia* are addressed. The first is whether short-read genomic data can be used to produce a reliable estimate of the size of a genome, even in the absence of an assembly, by comparing the estimated genome sizes with measurements obtained from flow cytometry. The second concerns the dynamics of the repetitive DNA fraction and the question of which repeat types contributed most to changes of genome size between species, and the pattern of dynamics of various repeats in the analyzed species.

To answer these questions NGS Illumina short-read paired-end data from species in the genus *Tillandsia*, and related genera, were analyzed. Firstly, two different methods were employed to estimate the genome size of 21 species (16 in *Tillandsia*, and 5 in the larger Tillandsioideae group), using short read genomic data. These are the exon alignment method, and the *k*-mer method. The results of both methods were compared to absolute genome sizes measured for all of the 21 species using flow cytometry.

Subsequently, the repeat compositions of the genomes of five selected *Tillandsia* species (*T. fasciculata*, *T. ionantha*, *T. leiboldiana*, *T. propagulifera*, and *T. australis*) were analysed using RepeatExplorer. These species of *Tillandsia* subgenus *Tillandsia* have been chosen because they are closely related, but differ in genome sizes (from 713 Mb in *T. juncea* to 1,112 Mb in *T. leiboldiana*). The analyses were performed for each species individually, and also in a comparative setting with all five taxa together. The repeat composition of the five species was analysed using RepeatExplorer. This allowed the characterization of the proportion of each genome composed of repeats, and the classification of the various repetitive elements. Comparison of the genome sizes and repeat composition of the five species allowed the inference of the relative repeat dynamics among the species.

2. Materials and Methods

2.1 Plant material

21 species were chosen for this project (Table 2). The plant material used for DNA extraction originated from collections in Botanical Gardens in Vienna, Heidelberg, Rio de Janeiro, Lyon, Porrentruy, and Geneva. NGS sequencing was performed with an Illumina HiSeq3000 Sequencing System, producing short, paired-end reads [de la Harpe et al., 2020; Table 3].

Table 2. Accession numbers of material analyzed in this study.

Species	Accession number (NGS data)	Accession no. of samples used in flow cytometry (where different)
T. australis *	HBV B138/87	
T. aff. australis	B1293 = HBV 0024576	
T. complanata	BRO 000726 Heidelberg	HBV 0029035
T. curvispica	B1597	HBV B393/96
T. demissa	HBV 0021466	
T. aff. divaricata	HBV 0024640	
T. fasciculata *	HBV 0024654	
T. floribunda	HBV 0029852 = BRO 140783	
T. ionantha *	HBV 0025215	
T. juncea	Tjunc_633a	HBV 0025397
T. latifolia	HBV 0023851	HBV 0024711
T. leiboldiana *	HBV 0000663 = BRO 000613	
T. propagulifera *	HBV 0023792	
T. rauhii	HBV 0024778	
T. somnians	PE-0-FRP-2821-FK3.2015	HBV 0024823 & HBV 0000087
T. sphaerocephala	HBV 0023378	
Barfussia laxissima	HBV 0024714	

Species	Accession number (NGS data)	Accession no. of samples used in flow cytometry (where different)
Lemeltonia dodsonii	HBV 0024642	
Pseudalcantarea macropetala	HBV 0025219	
Racinea spiculosa	HBV 0027886	
Wallisia lindeniana	HBV 0024851	

2.2 Analyses of Illumina DNA sequence data

Paired-end Illumina short-read DNA sequence data from 21 species in Tillandsioideae (16 in *Tillandsia*, and 5 from other genera) were used for genome size estimation (exon and *k*-mer methods) in order to evaluate the accuracy of these methods against known genome sizes measured using flow cytometry. Illumina sequence reads from five of the *Tillandsia* species were also used for repeat profile analyses with RepeatExplorer to determine the composition of the repetitive part of their genomes in a comparative context. Prior to the analyses, the sets of DNA sequence data were preprocessed to ensure data quality, to standardize the read lengths, and to remove non-nuclear DNA reads.

For all 21 species both short-read (Illumina) sequence data and flow cytometry measurements were available. For most species the sequence data and the measured genome size originated from the same individuals, with a few exceptions (Table 2). A single set of DNA sequence data was analyzed per sample. The data was provided as pairs of fastq files, with one file for forward reads and one for reverse.

Table 3. Statistics of the Illumina DNA sequence datasets.
The five species used for repetitive DNA analyses are indicated with *.

Species	Number of read pairs before preprocessing	Number of read pairs after preprocessing	Median read length before preprocessing
T. australis *	58,019,028	43,375,465	126
T. aff. australis	11,398,320	4,892,618	125
T. complanata	69,072,167	34,851,629	151
T. curvispica	10,839,939	4,743,802	125
T. demissa	10,664,792	4,748,017	125
T. aff. divaricata	10,262,400	4,582,652	125
T. fasciculata *	71,292,543	46,181,781	151
T. floribunda	64,357,854	43,509,036	151
T. ionantha *	10,841,869	4,791,130	125
T. juncea	41,516,860	31,108,339	151
T. latifolia	49,031,830	33,714,973	151
T. leiboldiana *	37,604,256	26,341,133	151
T. propagulifera *	69,767,081	48,072,357	151
T. rauhii	9,221,457	4,101,003	125
T. somnians	65,725,750	44,469,816	151
T. sphaerocephala	60,385,317	39,805,737	151
Barfussia laxissima	9,334,775	4,249,447	125
Lemeltonia dodsonii	9,797,604	4,267,705	125
Pseudalcantarea macropetala	9,727,880	3,701,708	125
Racinea spiculosa	8,447,581	3,560,884	125
Wallisia lindeniana	8,818,457	3,899,416	125

2.3 Illumina sequence data quality assessment and preprocessing

The quality of the sequence data was assessed using FastQC, version 0.11.7 [Andrews, 2010], which provided information about the quality score along the length of the reads, quality score across the tiles of the sequencer, and other metrics. The datasets had previously been filtered for quality, using ConDeTri, with base quality = 20 [de la Harpe et al., 2020] and were therefore expected to be of high quality. Most metrics indicated high quality sequences except for the 'Per Base Sequence Content' and the 'Sequence Length Distribution' for several samples. The sequence length of several samples was not uniform with majority of the reads being either 125, 126, or 151 bases long (Table 3). The data sets were therefore further filtered by trimming to obtain reads of identical length. Reads of all species were trimmed to 115 bases, removing low quality ends with trimmomatic, version 0.38 [Bolger et al., 2014], using the flags 'CROP:125 HEADCROP:10 MINLEN:115'. Subsequently, the reads that represented organellar DNAs (mitochondria and plastids) were filtered out using available organellar assemblies of *Ananas comosus* [de la Harpe et al., 2020]. Both trimming and removal of non-nuclear sequences were performed by the shell script "tillandsia_trim_&_chloro.sh". The *Tillandsia* data sets were aligned to *Ananas* mitochondrial and plastid sequences with bowtie2, version 2.3.5.1 [Langmead & Salzberg, 2012], and the reads which did align were discarded with samtools [Li et al., 2009]. Finally, bedtools [Quinlan & Hall, 2010] was used to convert the data back into fastq format. The trimmed and cleaned fastq files were subsequently used for three types of analysis: 1) the exon alignment method of genome size estimation; 2) the *k*-mer counting method of genome size estimation; and 3) the analysis of repeatomes with RepeatExplorer.

2.4 Genome size estimation I: the exon alignment method

This method uses the assumed homology between the exome of the analyzed *Tillandsia* samples and the exome of the related *Ananas comosus*, for which several genome assemblies exist, e.g. [Ming et al., 2015], to estimate the sequencing depth of the samples, and their 1C genome sizes. The method was implemented with the following software tools: bowtie2, samtools, and bedtools. In order to minimize errors from typos and for reproducibility, these packages were called from two shell scripts. First, "tillandsia_ananas_align.sh" used bowtie2, version 2.3.5.1, with the "--very-sensitive-local" flag, to align the short read data for each species to the *Ananas comosus* reference genome [Ming et al., 2015]. The resulting sam file was converted to a bam file, using samtools, version 1.8, and duplicated and low quality entries were removed. The final output file contained an alignment of the *Tillandsia* DNA sequence data for each species to the *Ananas* reference assembly.

The second script, "tillandsia_indexing_&_exon_coverage.sh" uses samtools to build an index for each of the alignment files, and bedtools, version 2.27.1, (with the "bedtools coverage" command) to find the depth of coverage for every exon in each of the analyzed species. The final output of the two scripts was a single number for each set of DNA sequence data, giving an estimate of the average depth of coverage. To obtain the estimated 1C genome size, the total length of reads (read length times number of reads) was divided by the estimated coverage.

2.5 Genome size estimation II: the *k*-mer method

The first step in this method is to count the *k*-mers present in the genome, in order to produce the *k*-mer frequency spectrum. This was performed using jellyfish, version 2.2.8 [Marçais & Kingsford, 2011] with two commands: (1) "jellyfish count", which takes a fastq file as its input and creates a temporary file that contains all the *k*-mers of the specified length, counting each *k*-mer and its reverse complement

together; and (2) "jellyfish histo" counts the *k*-mers in the temporary file, and outputs a file giving the number of *k*-mers with each multiplicity, which can be plotted as the *k*-mer frequency spectrum. Several values of *k* were tested: *k* = 13, 15, 17, and 19. The value *k* = 13 was found to be best for most of the samples with the resulting peaks in the histogram being clearly identifiable in 10 species (for example Figure 3a), but not in the remaining 11 (Figure 3b). The histograms were visualized in Rstudio. An attempt to use an R package called findGSE [Sun et al., 2017] to locate the peaks was unsuccessful, so the positions of the peaks were determined by eye.

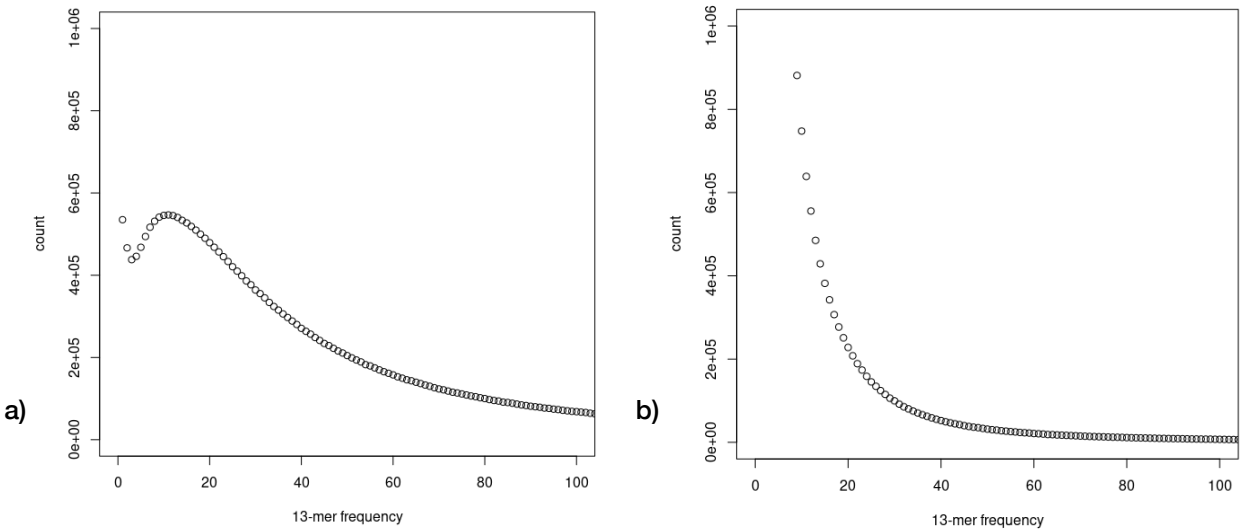


Figure 3. *k*-mer frequency plots.
a) *T. fasciculata* (exon method coverage = 7.05x)
b) *T. ionanatha* (exon method coverage = 0.87x)

2.6 Repeatome analysis

RepeatExplorer was used to identify and quantify the repeat families and types present in the genomes of five of the *Tillandsia* species. The software uses as its input short-read DNA sequence data, and is therefore ideal for investigating species for which there is no genome assembly available. The algorithm of RepeatExplorer performs an all-to-all comparison of reads, with a threshold of 90% sequence similarity over at least 55% of the read length (and at least 55 nt). Sequence similarity between the reads is represented by a graph where each read is a node and each pair

of reads which are sufficiently similar are connected by an edge. The structure of the graph is then examined to detect clusters of frequently connected nodes. Each cluster represents one type of repeat or part of a repeat. The use of paired end reads allows the grouping of clusters into superclusters. Two clusters are placed in the same supercluster if there is a sufficient number of read pairs which have one read in each of the two clusters [Novák et al., 2020]. In addition to detecting repeats by clustering, RepeatExplorer also provides an initial annotation of the clusters and superclusters. The reads in each cluster are assembled into pseudo-consensus contigs, using CAP3 [Huang & Madan, 1999; Novák et al., 2020]. These contigs are BLASTed against a library of common repeats, REXdb [Neumann et al., 2019]. Identification of tandem repeats is also implemented in the pipeline with TAREAN [Novák et al., 2017; Novák et al., 2020]. The results of this automatic annotation have to be improved by subsequent manual annotation. The RepeatExplorer pipeline was run on the Czech Galaxy server, which is hosted at https://repeatexplorer-elixir.cerit-sc.cz/galaxy [Chudoba et al., 2017; Novák et al., 2020]. RepeatExplorer was used to analyze the repeatome of five of the *Tillandsia* species - *T. australis*, *T. fasciculata*, *T. ionantha*, *T. leiboldiana*, and *T. propagulifera*. First, the analysis was performed individually, and then the reads from individual species analyses were used to perform a comparative analysis of all five species together, in order to quantify the differences in repeat composition between the species.

2.6.1 Individual analyses

The five pairs of fastq files (forward and reverse reads) were uploaded to the RepeatExplorer website. For each species, the forward and reverse reads were interlaced into a single file, using the 'Preprocessing of FASTQ paired-end reads' tool and then converted from fastq to fasta using the converter tool. The interlaced file was submitted to the RepeatExplorer clustering tool, selecting the 'paired-end reads' and 'long queue' options, allowing more exhaustive analyses to be run. After each run was finished, a large number of files was available for download, among them

csv tables of the annotated clusters of repeats, repeat abundances, and images of the graphs of all clusters and superclusters, showing the protein domains of the elements that had been identified by automatic annotation. The files were downloaded to a local computer as zip archives, so that the information could be used as the starting point for further manual analysis of the results.

RepeatExplorer classifies the clusters into either "top" clusters (containing more than 0.01% of the input reads) or "small" clusters. Only top clusters are annotated. Manual annotation allows for more accurate identification of different types of repeats that differ in their biology and genetic structure. This is particularly important for the correct annotation of non-coding satellite DNAs. Dotter [Sonnhammer & Durbin, 1996] was used to correct for frequently false-positive annotation of such tandem repeats and additionally also to annotate other repetitive elements, such as MITEs and Cassandra-TRIMs, that can be identified only using their structural features. Manual annotation allows for informed decisions on the classification of chimeric superclusters that may be composed from more than one repeat type. Finally, clusters corresponding to contamination (e.g., remnants of organellar DNAs, spike-in DNAs) are also identified and so can be excluded from the final annotation.

2.6.2 Comparative analysis

In addition to detecting and annotating repeats in a set of reads from an individual sample, RepeatExplorer is also able to perform a comparative analysis, in which reads from multiple samples are combined into a single fasta file and analysed together, so that the repetitive DNA content of several species can be directly compared. This allows the identification of particular repeat lineages whose abundance differs between the species.

Comparative analysis is typically done following individual analysis of each sample, by concatenating read files from single species analyses and using this as the input file for the comparative analysis. To create the input for the comparative

analysis all of the clustered (unpaired) reads from an individual analysis were concatenated into a single file. A shell script ("prepare_reads_comparative.sh") was used to "re-pair" the incomplete pairs and to order the reads so that they were properly interlaced. Prefixes were added to all the read names to identify the species they represented. The commands 'seqtk seq -1' and 'seqtk seq -2' were used to separate each file into two files, one with forward and one with reverse reads. Subsequently, 'seqtk sample -s123' was used to sample the reads, using the same sample seed to make sure the corresponding half-pairs are chosen from each of the two files. An equal number of reads was sampled from all five species (500,000 pairs each). The command 'seqtk mergepe' was used to merge and interlace the two files. Finally, the five separate fasta files were then concatenated into one. This file was uploaded to Galaxy, and submitted to RepeatExplorer clustering as the input for the comparative analysis.

2.6.3 Improvement of annotations by comparison of individual and comparative analyses

The annotation of comparative analysis clustering was improved by comparing the annotation of the reads in the comparative analysis with the annotation of the clusters into which these reads were placed in the individual analyses. The script "cluster_comp_by_read.sh" was used to compare the lists of read names in the two sets of output data (comparative and individual), which produced a text file listing the corresponding cluster numbers in the two sets of results, and additionally the proportions of reads that the corresponding clusters have in common. This information, building on the automatic annotation of the comparative analysis, aided the final round of manual annotation of the comparative analysis.

In order to implement the methods used, some short pieces of code were written. All the code written for this project is available at https://github.com/neilmcnair/tillandsia .

3 Results

3.1 Estimation of genome sizes

The depth of genome coverage was estimated for each of the 21 NGS paired-end short-read datasets, with the two methods (Table 4), and estimated genome sizes were derived from the depths of coverage. The *k*-mer method resulted in estimates for only 10 of the datasets, for all of which the estimates from *k*-mer method were higher than those from the exon method.

Table 4. Depths of genome coverage and 1C genome sizes of the 21 analyzed species.

All genome sizes are in Mb. (†) indicates the mean of two measurements.

Species	Measured genome size (flow cytometry)	Exon method		*k*-mer method	
		Estimated coverage	Estimated genome size	Estimated coverage	Estimated genome size
T. australis *	935.4	4.97	2,007	7.26	1,375
T. aff. australis	964.1	0.72	1,561		
T. complanata	857.8	6.60	1,215	7.82	1,026
T. curvispica	917.0	1.24	877		
T. demissa	876.4	0.79	1,391		
T. aff. divaricata	939.2	0.67	1,567		
T. fasciculata *	797.7 (†)	7.05	1,506	12.28	865
T. floribunda	896.0	5.25	1,908	10.05	996
T. ionantha *	736.5	0.87	1,263		
T. juncea	712.7	7.13	1,004	11.17	641
T. latifolia	976.7	4.11	1,888	7.26	1,069
T. leiboldiana *	1,111.8	1.85	3,273	4.47	1,357
T. propagulifera *	885.0	5.59	1,978	11.17	990

Species	Measured genome size (flow cytometry)	Exon method		*k*-mer method	
		Estimated coverage	Estimated genome size	Estimated coverage	Estimated genome size
T. rauhii	744.1	0.86	1,099		
T. somnians	711.3 (†)	7.70	1,328	11.17	916
T. sphaerocephala	1,265.5	5.70	1,606	7.82	1,171
Barfussia laxissima	943.1	0.58	1,693		
Lemeltonia dodsonii	988.9	0.60	1,635		
Pseudalcantarea macropetala	397.5	1.33	640		
Racinea spiculosa	954.4	0.43	1,892		
Wallisia lindeniana	874.7	0.56	1,595		

The estimates were compared to the absolute genome sizes measured for all 21 species using flow cytometry. All but one of the measured genome sizes were larger than that of *A. comosus*, with 591 Mb [Chen et al., 2019]. For 19 of the 21 species, the genome size estimates from the exon alignment method were larger than the measured values; these estimates were on average 77% higher (ranging from +194% to -4%). For *T. curvispica* the estimated value was smaller than the measured value. For *T. juncea* the estimated value was between the two measured values.

The *k*-mer method resulted in a genome size estimate for only ten of the 21 species; these estimates were on average 14% higher than genome sizes measured using flow cytometry (ranging between +47% and -10%). For eight of the ten species, the estimates were higher than the measured genome sizes. For *T. juncea* and *T. sphaerocephala* the estimates were lower. The value of *k* that was used was 13.

3.2 Individual analyses of repeatomes

The analysis of the repeatomes of the five chosen species (*T. australis, T. fasciculata, T. ionantha, T. leiboldiana,* and *T. propagulifera*) showed that all five have a high abundance of genomic repeats, ranging from 81.1% to 85.5%. The species differ in the abundances of the different types of repeats. The results of the final annotation allowed the inference of the proportions of the various repeat types in the analyzed genomes (Table 5).

Table 5. Composition of repetitive DNA fractions expressed as proportion of each genome (in %).

		T. australis	*T. fasciculata*	*T. ionantha*	*T. leiboldiana*	*T. propagulifera*
Mobile elements - Class I						
LTR/Ty1-copia	SIRE	18.76	16.51	14.28	21.26	9.55
	other copia[1]	2.31	2.45	2.54	1.71	2.19
LTR/Ty3-gypsy	Tekay	39.27	37.06	37.90	47.40	50.72
	other gypsy[2]	1.24	0.31	0.26	0.12	0.46
LTR/Cassandra-TRIM		0.26	0.31	0.32	0.51	0.20
LTR/unclassified		1.16	0.64	0.62	0.58	1.38
Non-LTR	LINE	—	0.05	0.04	0.03	—
Mobile elements - Class II						
Subclass I	TIR/ MITE	1.33	0.73	0.61	0.20	0.69
	other TIR[3]	1.06	1.11	1.29	1.28	1.09
Subclass II	Helitron	0.14	0.09	0.36	0.07	0.08
"mixed" clusters - Class I & Class II	Tork & hAT	0.15	0.22	0.24	0.19	0.18

		T. australis	*T. fasciculata*	*T. ionantha*	*T. leiboldiana*	*T. propagulifera*
Tandem repeats						
rDNA	35S	0.42	0.28	2.07	0.38	0.47
	5S	0.26	0.24	0.96	0.26	0.08
Satellite DNAs		0.06	0.16	0.15	—	0.03
Unclassified repeats		5.42	5.31	3.93	2.43	4.52
Total top clusters		71.85	65.46	65.59	76.42	71.63
Small clusters		13.04	15.60	17.13	9.05	12.82
Singletons		15.11	18.93	17.28	14.53	15.55

[1]"other *copia*" consists of Ale, Alesia, Angela, Ikeros, Ivana, TAR, and Tork; [2]"other *gypsy*" consists of CRM, Galadriel, Ogre, and Reina; [3]"other TIR" consists of EnSPm_CACTA, hAT, MuDR_Mutator, and PIF_Harbinger.

In each of the five genomes, LTR retrotransposons were the most abundant repeat type, representing from 55.6% of the genome in *T. ionantha* (Figure 4c) to 71.1% in *T. leiboldiana* (Figure 4d). Among LTR retrotransposons, Ty3-*gypsy*/Tekay (Figure 6a) and Ty1-*copia*/SIRE (Figure 6b) families were most abundant, and together constituted more than 93% of all LTR retrotransposons. Tekay had abundances of between 37.1% (in *T. fasciculata*) and 50.7% (in *T. propagulifera*), and in all five species was more abundant than SIRE, which had abundances of between 9.6% (in *T. propagulifera*) and 21.3% (in *T. leiboldiana*). In four of the species, the ratio of the abundance of Tekay to SIRE ranged from 2.1 to 2.7, while in *T. propagulifera* (Figure 4e), the ratio was 5.3, corresponding to both the highest abundance of Tekay and the lowest abundance of SIRE.

Other types of LTR retrotransposon occurred at much lower abundances. The Ty1-*copia* families of Ale (Figure 6c), Alesia, Angela, Ikeros, Ivana, TAR, and Tork each comprised less than 0.7% of each genome. The Ty3-*gypsy* families of CRM, Galadriel, and Reina each comprised less than 0.4% of each genome. Ty3-*gypsy*/Ogre had an abundance of 1% in *T. australis*, but much lower abundances in other species. Cassandra-TRIM, a non-autonomous element which consists of a piece of 5S rDNA

inside an LTR (Figure 7c), was found in all five species with abundances of between 0.2% and 0.5%. Non-LTR retrotransposons (LINEs) were found in three species (*T. fasciculata*, *T. ionantha*, and *T. leiboldiana*) with very low abundances (0.03% - 0.05%). Class II transposable elements, of both subclass I (TIR) and subclass II (Helitron) were present in lower copy numbers than class I elements. The TIR elements were represented by EnSpm_CACTA, hAT, MuDR_Mutator, PIF_Hardinger, and MITEs, all with abundances of 1.3% or less. Helitrons were found in all five species with abundances between 0.07% and 0.36%. One cluster containing reads from both class I and class II elements, namely Tork (class I/Ty1-*copia*) and hAT (class II/TIR), was consistently identified in all five species (Figure 6d), with abundances between 0.15% and 0.24%.

For each of the five species, the abundances of genomic elements were visualized as a hierarchical pie chart (Figure 4), plotted using Krona, which is available on the Galaxy website. Also shown are repeat abundances (Figure 5), where small clusters and singletons have been excluded. RepeatExplorer graphs were used to aid identification of different repeat types (Figures 6 & 7). Repeats were usually identified from single clusters (Figures 6a, 6c, 6d, 7b). There were also cases where the repeat type was identified from protein domains distributed among several clusters joined in a supercluster (Figures 6b, 7a).

Tandem repeats represented only a very small proportion of the analyzed genomes and included both rDNAs and a putative novel satellite DNA with monomer length between ~1300 and ~1600, which was identified in four species (*T. australis*, *T. fasciculata*, *T. ionantha*, and *T. propagulifera*) albeit with very low abundances (0.03% - 0.16%). rDNAs were recovered in all five species, also in low abundances (0.5% - 3.0%). MITEs and satellite DNAs have similar graph structure, but can be identified by alignment visualization tools, such as Dotter (Figure 8).

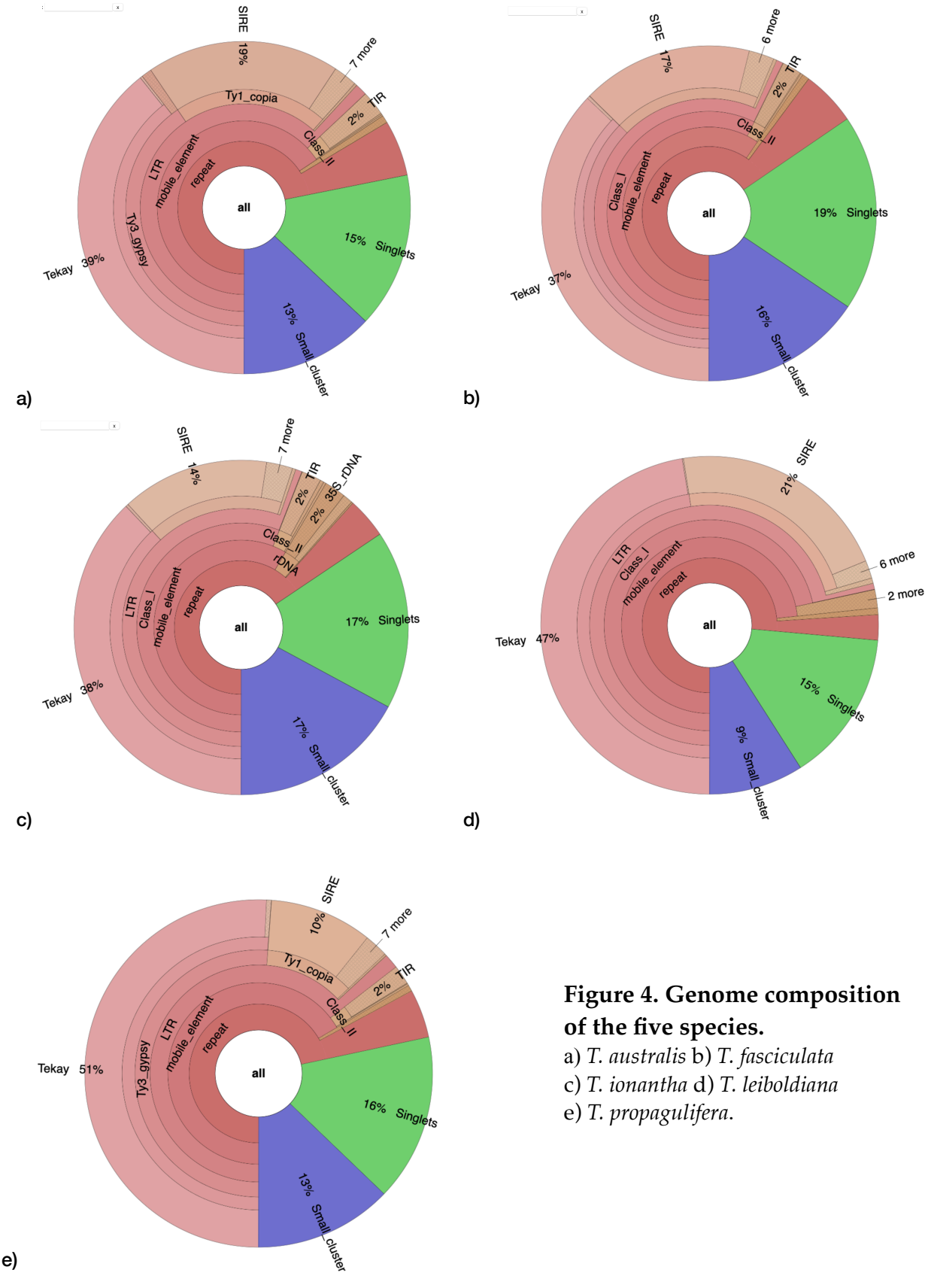


Figure 4. Genome composition of the five species.
a) *T. australis* b) *T. fasciculata* c) *T. ionantha* d) *T. leiboldiana* e) *T. propagulifera*.

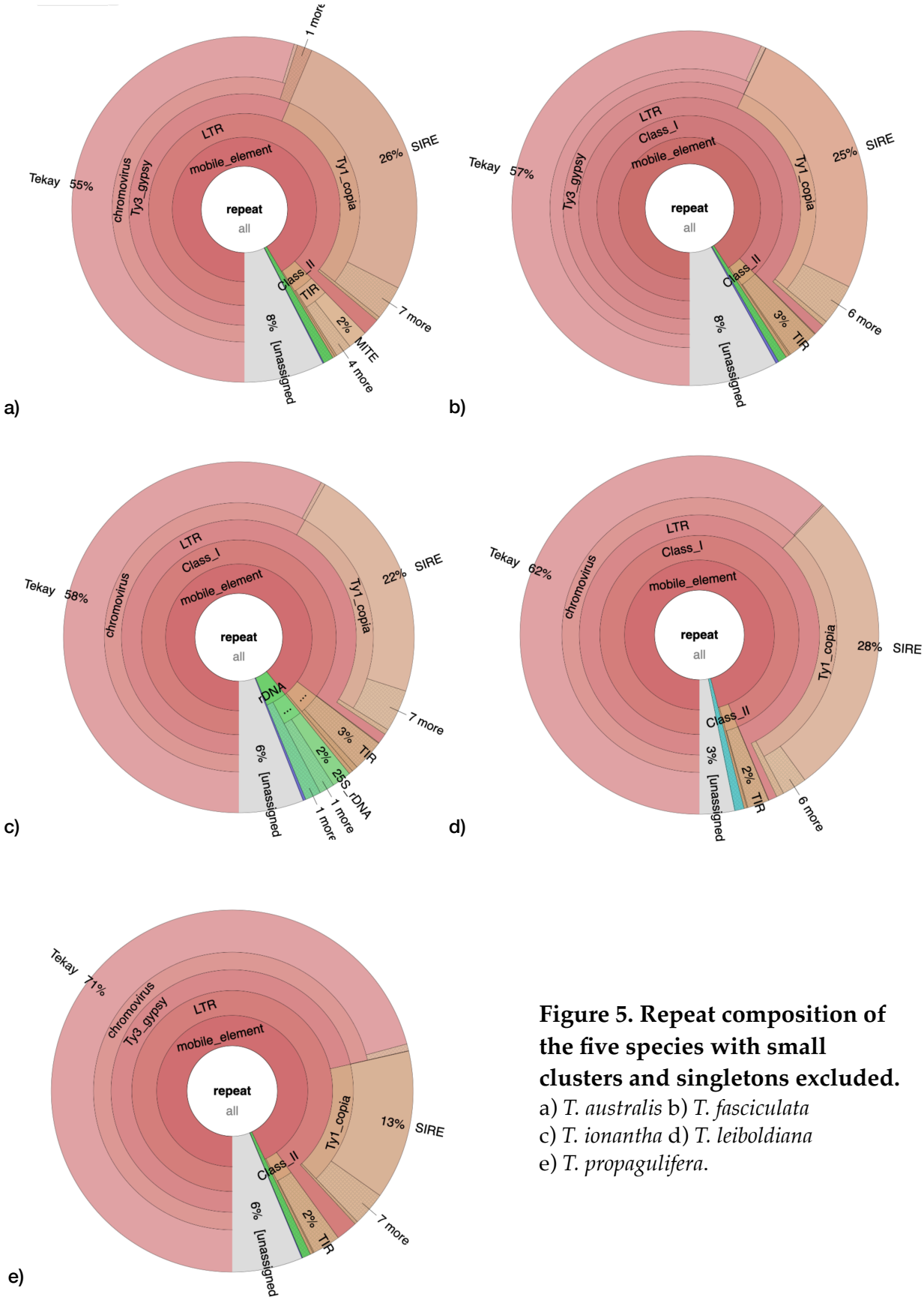


Figure 5. Repeat composition of the five species with small clusters and singletons excluded.
a) *T. australis* b) *T. fasciculata*
c) *T. ionantha* d) *T. leiboldiana*
e) *T. propagulifera*.

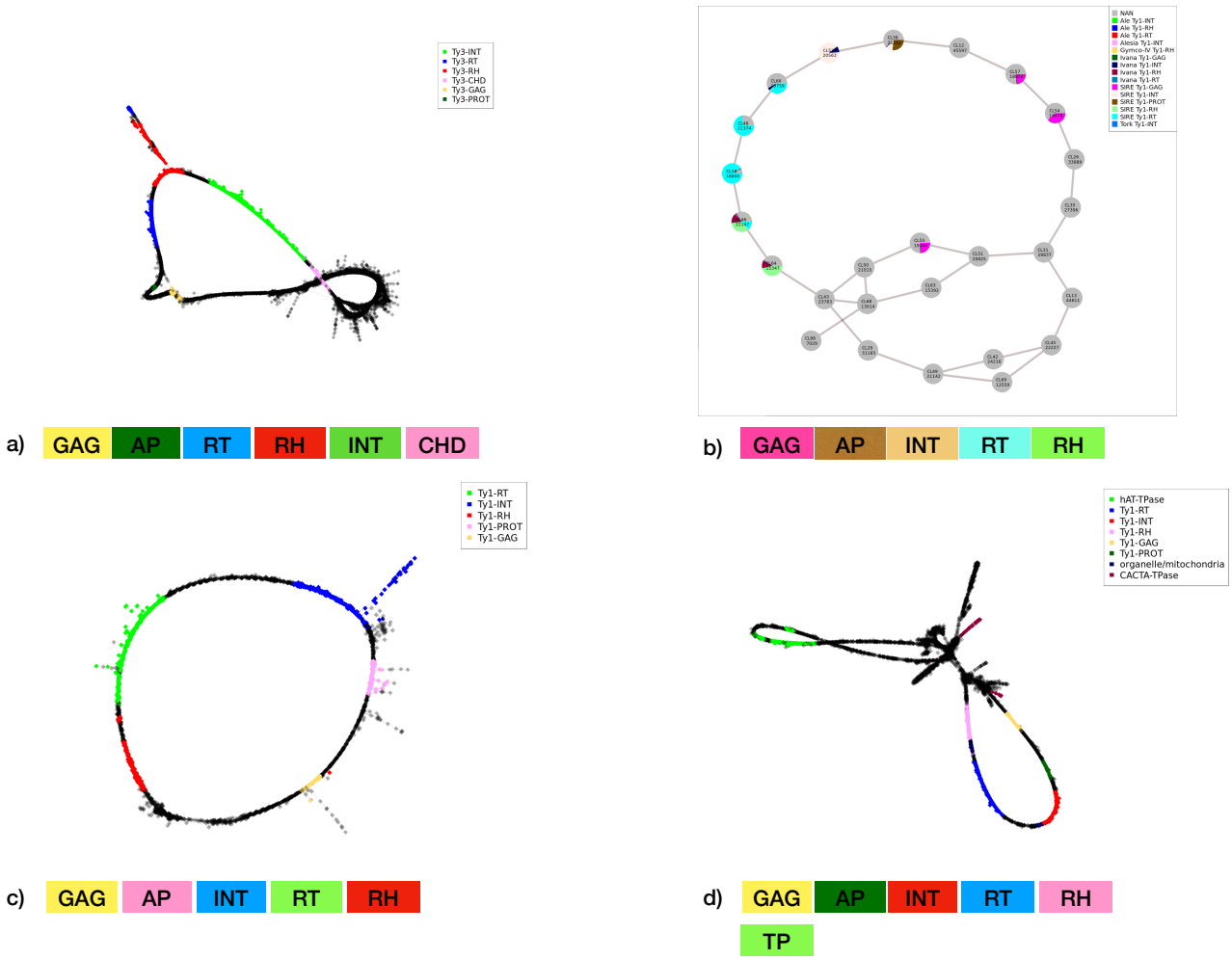


Figure 6. RepeatExplorer graphs of different transposon types recovered in comparative analysis.

a) Ty3-*gypsy* / Tekay (cluster CL65); b) Ty1-*copia* / SIRE (supercluster SC2); c) Ty-*copia* / Ale (cluster CL123); d) a "mixed" cluster, containing reads from both Ty1-*copia* / Tork and TIR / hAT (cluster CL87).

Abbreviations: AP - (aspartic) protease; CHD - chromodomain; GAG - group specific antigen; INT - integrase; RH - ribonuclease H; RT - reverse transcriptase; TP - transposase.

The lowest abundances of unique reads (singlets) and of "small" clusters (with fewer than 0.01% of total reads) both occurred in *T. leiboldiana* (15% and 9% of the genome, respectively). The highest abundance of singlets was found in *T. fasciculata* (19%), and the highest abundance of small clusters was found in *T. ionantha* (17%). The proportion of each genome in top clusters varied from 65% in *T. fasciculata* to 76% in *T. leiboldiana* (Figure 4).

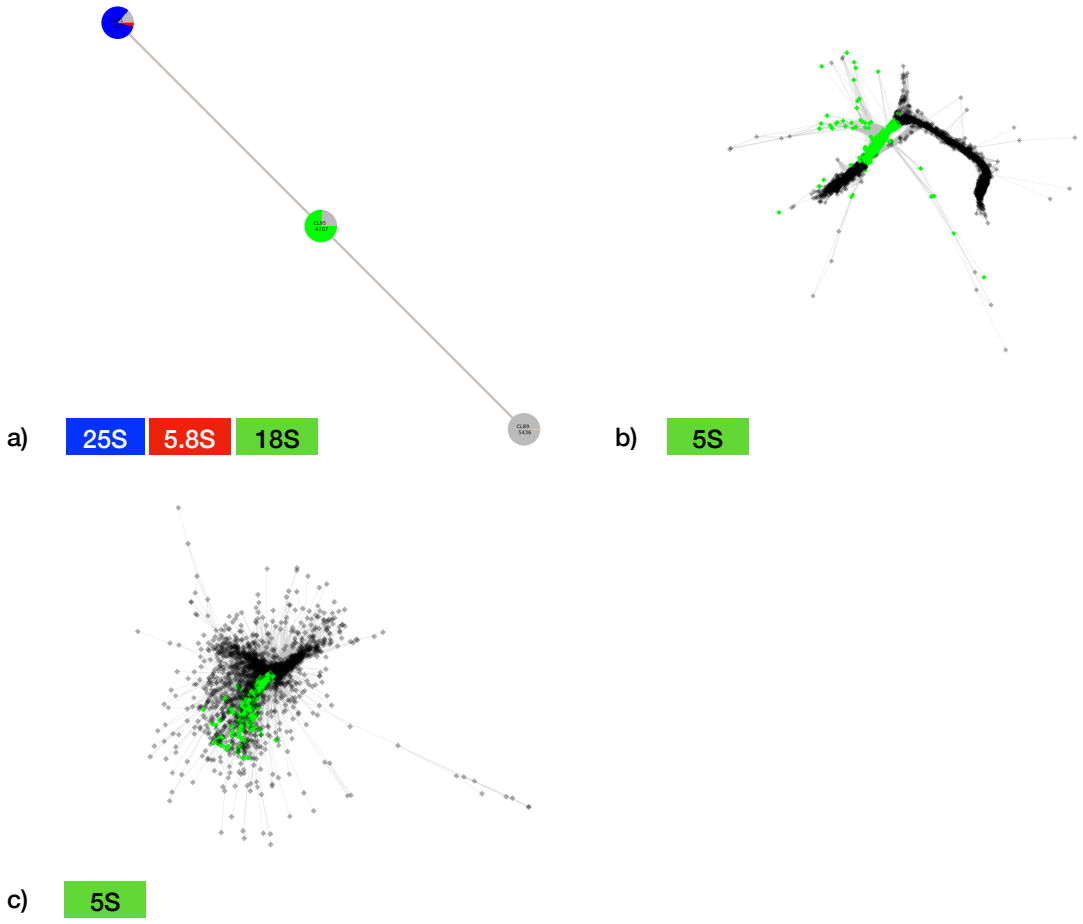


Figure 7. RepeatExplorer graphs of rDNA, and Cassandra-TRIM.
a) 35S rDNA (supercluster SC9 in *T. australis*); b) 5S rDNA (cluster CL84 in the comparative analysis); c) Cassandra-TRIM (cluster CL75 in the comparative analysis)

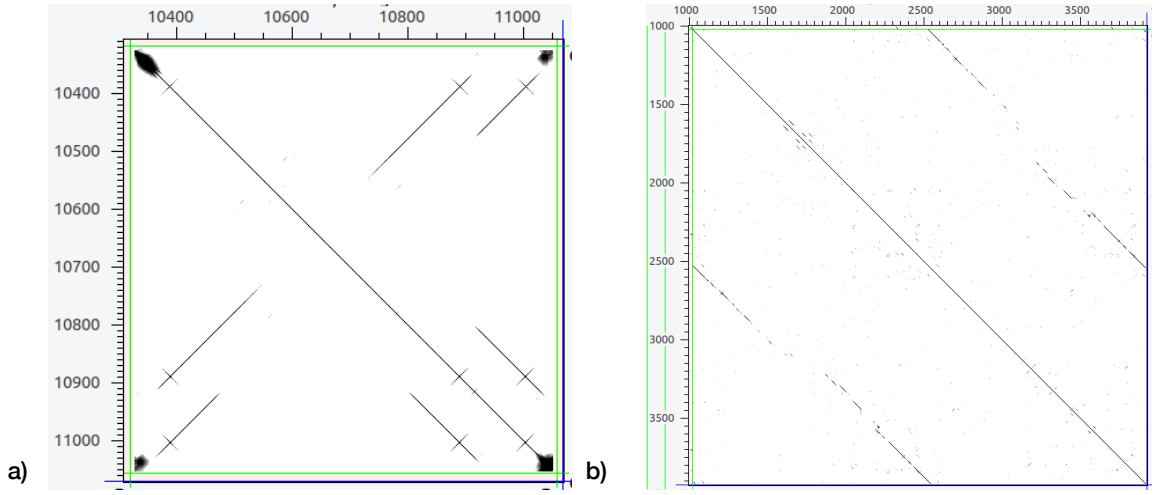


Figure 8. Dotter plots of a MITE and of a satellite DNA recovered in the comparative analysis.
a) MITE (cluster CL141) showing the characteristic "X" shape; b) Satellite DNA (cluster CL181) showing parallel diagonal lines and a monomer length of between ~1300 and ~1600 bp.

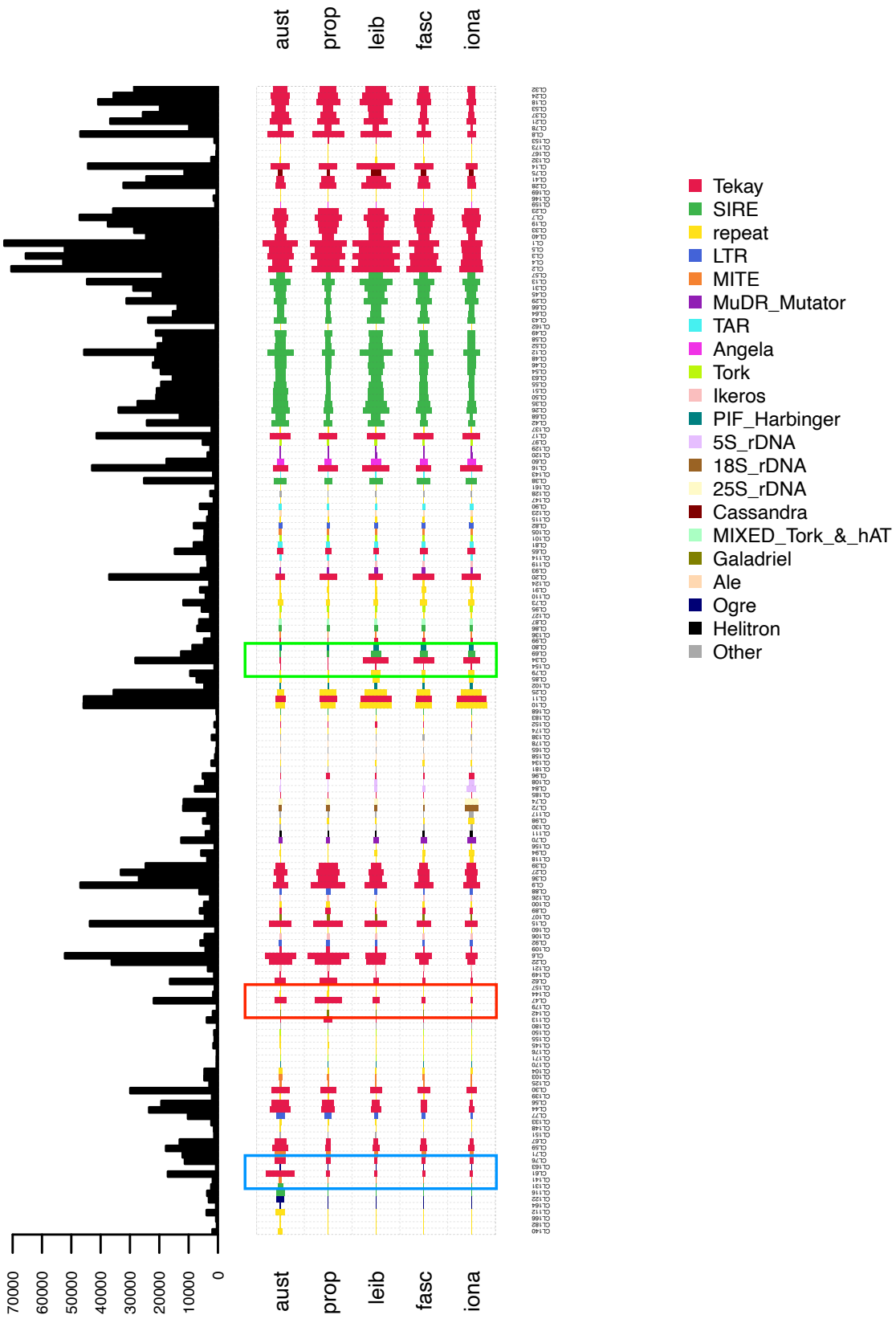
3.3 Comparative analysis of repeatomes

In order to better describe the differences between the species, a comparative analysis was performed, pooling the clustered reads from the individual analyses of the five species to determine the relative abundances of each cluster across the five species, calibrated using measured genome sizes (Figure 9). These repeat profiles indicated high levels of similarity between *T. fasciculata*, *T. ionantha*, and *T. leiboldiana*, while *T. australis* and *T propagulifera* were more distinct from the other three species and also from each other.

All five species had large abundances of Ty3-*gypsy*/Tekay but each of the five repeatomes exhibited differences, mostly in the proliferation profiles of different lineages of this element in different species. Comparative analysis allowed the identification of these distinct lineages. Cluster CL1 was much more abundant in *T. australis* than in the other species (blue rectangle in Figure 9), CL47 much much more abundant in *T. propagulifera* than in the other species (red rectangle), and CL34 much more abundant in *T. fasciculata*, *T. ionantha*, and *T. leiboldiana* than in the other two species (green rectangle).

Figure 9. Comparative analysis of the five species.

Each row is a cluster of reads, with the class of repeat indicated by colour, and the relative abundances by the lengths of the blocks.
Abbreviations: aust - *T. australis*; prop - *T. propagulifera*; leib - *T. leiboldiana*; fasc - *T. fasciculata*; iona - *T. ionantha*.



4. Discussion

This project addressed the question of how accurately genome sizes could be estimated using only short read DNA sequence data, in selected species in *Tillandsia*. The second issue addressed was the composition of the repeatomes of five of the species, and how well the genome size evolution of these species could be understood by reference to the differences between the repeatomes.

4.1 Efficacy of NGS-data based genome size estimation methods

Both genome size estimation methods produced estimates that were generally higher than the measured values. The exon method assumes a high level of similarity between the genomes of the study species and the genome of the reference species. However *Ananas comosus* may not be sufficiently closely related to *Tillandsia* to allow for accurate estimation of genome sizes. The divergence time of Tillandsioideae from Bromelioideae (the lineage containing *Ananas*) has been estimated to be 15.4 Mya [Givnish et al., 2011]. If *Tillandsia* and *Ananas* are too distantly related, then the alignment between their exonic regions will be less accurate, and this will result in a shorter estimate of the length of the exonic region of each *Tillandsia* (or Tillandsioideae) genome. The estimated depth of coverage will be too low and therefore the estimate of the genome size will be too high. As this was observed in all *Tillandsia* species but one, it can be concluded that *A. comosus* may be too distant to be a suitable reference genome for the exon method. The best approach to improve the genome size estimates from the exon alignment method would be to use a genome assembly from a phylogenetically closer taxon. Such assemblies have recently become available, for species in *Tillandsia* itself [Groot Crego et al., 2024].

Genome size estimates from the *k*-mer method were obtained for only 10 species, with eight of them inferred to be larger than the measured values. Both the

failure to obtain estimates and the too-large estimates of genome size that were obtained can be explained by the depth of coverage of the datasets. Using the coverage estimates from the exon method as a comparison (so that figures are available for all 21 species) it is seen that the species for which the k-mer method produced estimates were the 10 species with the highest depths of coverage (between 1.85x and 7.7x); while the remaining species had lower coverages (between 0.43x and 1.33x). *T. fasciculata* had a coverage of 7x (by the exon method) and this was sufficient to produce a *k*-mer frequency plot with a clear peak (Figure 3a). In contrast, *T. ionantha* with a coverage of less than 1x (by the exon method) did not result in a *k*-mer frequency plot with a clearly defined peak and thus the method failed to provide an estimate (Figure 3b). The recommended sequencing depth has previously been discussed as an important factor in the *k*-mer method [Hesse, 2023]. A coverage of approximately 1x is been deemed insufficient to distinguish the main peak from the peak at frequency = 1, which represents sequencing errors, and thus no estimate can be obtained. At coverages of approximately 5x, the main peak is distinguishable, but its shape is distorted by proximity to the error peak, leading to an underestimate of the *k*-mer coverage, and therefore to overestimation of the genome size. This is seen in this study for the 10 species for which estimates were obtained.

The value of *k* that was used was 13. Higher values (15, 17) were tested but produced curves with peaks that were even less distinct. An attempt was made to use an R package findGSE [Sun et al., 2017] to locate the peaks, but was not successful. For almost all samples, the package failed to produce an estimate of peak position. The performance of findGSE was previously successfully tested on samples with coverages above 20x [Sun et al., 2017], while the highest coverage of species in this study was only 7.7x. In addition, findGSE requires a minimum *k*-mer length of 15, for which fewer of the graphs had clear peaks in the *Tillandsia* datasets. The best approach to improve the genome size estimates using the *k*-mer method would be to obtain sets of sequence data with higher depths of coverage. A value of approximately 20x or higher would be recommended [Hesse, 2023].

Comparison of the genome size estimates obtained from the two methods leads to the conclusion that the *k*-mer method estimates were closer to the measured values. However, given that the accuracy of the results was limited by the need for a closer reference genome assembly (for the exon method) and for datasets with greater depth of coverage (for the *k*-mer method), a direct comparison of the performance of two methods cannot be made. The new *Tillandsia* reference genomes [Groot Crego et al., 2024] should allow for the improvement of the results obtained using the exon method, while improvements in the *k*-mer method results would require DNA sequence datasets with a depth of coverage of 20x or more.

4.2 Evolutionary dynamics of the repeatome in selected *Tillandsia* species

The similarity of the abundances of different repeat lineages among the five analyzed *Tillandsia* species corresponded well to the phylogenetic relationships of these species [Barfuss et al., 2016; de la Harpe et al., 2020], with *Tillandsia fasciculata*, *T. ionantha*, and *T. leiboldiana* closely related to each other in clade K, *T. propagulifera* is more distantly related, but still part of *T. subg. Tilllandsia*, and *T. australis* more distantly related to the other four. Repeat profiles reflect whole genome architecture and have therefore been previously used for phylogenetic inferences [Dodsworth et al., 2015]. This phylogenetic signal is often obscured when considering other phenotypic characters. One such character in *Tillandsia* is genome size, with *T. leiboldiana* and *T. ionantha* having the largest and smallest measured genome sizes of the five species despite being closely related, and another is photosynthetic type, with *T. leiboldiana* and *T. australis* both using C3 despite not being closely related while the other three species use CAM.

Total transposon abundances ranged between 58% and 73% in the five analyzed species. In plant genomes analyzed to date abundances range from 3% to 85% [Mhiri et al., 2022] and these *Tillandsia* species are in the higher part of the range. In *Ananas comosus* the abundance of TEs has been found to be 70% [Chen et al., 2019] and the

peak age of LTR-retrotransposons in *A. comosus* was estimated to be ~1.8 Mya, long after the divergence of Tillandsioideae and Bromelioideae (~15.4 Mya), implying that at least some of the LTR-retrotransposons found in *Tillandsia* proliferated after the divergence of these genera. In the comparative analysis performed in this thesis, lineages of Ty3-*gypsy*/Tekay were found in differing abundances in the five species. The results suggest that proliferation of LTR-retrotransposons has been occurring in *Tillandsia* both after the divergence from Bromelioidaceae, as well as more recently, after the divergence between *T. subg. Tillandsia* and the *T. australis* complex, and also after the later divergence of the Mexican clade and the rest of *T. subg. Tillandsia* [Barfuss et al., 2016]. Analyses of more *Tillandsia* species will allow more precise estimation of the timing of bursts of transposable element proliferation.

Within the Mexican clade, *T. leiboldiana* had the largest genome size, as well as the highest abundances of Ty3-*gypsy*/Tekay and Ty1-*copia*/SIRE. The comparative analysis showed that many lineages of both these elements were more abundant in the repeatome of *T. leiboldiana* than in those of the other two species, *T. fasciculata* and *T. ionanatha*. The chromosome number of *T. leiboldiana* has been found to be $2n = 38$, while *T. fasciculata* had $2n = 50$, the common value for most *Tillandsia* species that have been reported [Groot Crego et al., 2024]. This suggests that the increase in the genome size of *T. leiboldiana* may be linked both to the proliferation of multiple lineages of repeats and to chromosomal rearrangements that have likely resulted in descending dysploidy.

The abundance of DNA transposons in all five *Tillandsia* species was less than 2% and therefore much lower than in *A. comosus* at 22% [Chen et al., 2019]. In whole genome assemblies of *T. fasciculata* and *T. leiboldiana*, annotated using the EDTA pipeline, DNA transposons were found with abundances of 3.5% and 9.6% respectively, compared to only 1.9% and 1.5% found in this project [Groot Crego et al., 2024]. These differences may be due to EDTA and RepeatExplorer using different libraries of repetitive element sequences and different algorithms to identify them. However it is also likely that the difference in sequencing technology may make a difference, specifically the use of long read sequence data versus the short read

Illumina sequence data used here. Although theoretically unbiased, in reality this may result in some technical bias in repeat content. Families of repeats that are especially GC or AT rich and therefore more fragile may be underrepresented in the datasets due to DNA fragment length selection (typically 500-800 bp) prior to sequencing, as has been shown for satellite DNAs [Emadzade et al., 2014].

The proportion of unclustered reads, or singletons, in each species corresponds to the quantity of non-repetitive genomic material. These values of the putative coding fraction of the genomes ranged from 127 Mb in *T. ionantha* to 162 Mb in *T. leiboldiana*. There was less variation in the size of this part of the genome than in the total genome sizes, as expected. The coefficient of variation among the five species was 0.16 for the total genome size, and only 0.09 for the size of the non-repetitive, putatively coding component. This supports the hypothesis that genome size variation between related species is caused by changes in the size of the repeatome rather than by variation in the size of the non-repetitive component. This can also been seen by considering the ranking of the species by genome size, and by repeat abundance. Among the five species studied, the three with largest genomes are *T. leiboldiana*, *T. australis*, and *T. propagulifera*. These three species, in the same order, also harbour the highest repeat abundances, again confirming that genome size changes are linked to repeat abundances.

The five species studied here are only a small proportion of the ~600 species in *Tillandsia*, and there is ample scope to use the techniques from this project to further investigate the dynamics of repetitive genomic elements, and to increase our understanding of the evolutionary history of this most interesting genus.

5. References

Andrews S. (2010) FastQC 0.11.7: a quality control tool for high throughput sequence data. (downloaded 2018-05-01 from http://www.bioinformatics.babraham.ac.uk/projects/fastqc)

Baranyi M., Greilhuber J., Święcicki W.K. (1996) Genome size in wild *Pisum* species. *Theor. Appl. Genet.* **93**: 717—721.

Barfuss M.H.J., Samuel R., Till W., Stuessy T.F. (2005) Phylogenetic Relationships in subfamily Tillandsioideae (Bromeliaceae) Based on DNA Sequence Data from Seven Plastid Regions. *Am. J. Bot.* **92**: 337—351.

Barfuss M.H.J., Till W., Leme E.M.C., Pinzón J.P., Manzanares J.M., Halbritter H., Samuel R., Brown G.K. (2016) Taxonomic revision of Bromeliaceae subfam. Tillandsioideae based on a multi-locus DNA sequence phylogeny and morphology. *Phytotaxa* **279**: 1—97.

Bennetzen J.L., Ma J., Devos K.M. (2005) Mechanism of recent genome size variation in flowering plants. *Ann. Bot.* **95**: 127—132.

Bennetzen J.L., Wang H. (2014) The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu. Rev. Plant Biol.* **65**: 505—530.

Benzing, D.H. (2000) *Bromeliaceae: Profile of an Adaptive Radiation*. Cambridge University Press, Cambridge, United Kingdom

Bolger A.M., Lohse M., Usadel B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114—2120.

Brown J.K., Gilmartin A.J. (1989) Chromosome numbers in Bromeliaceae. *Am. J. Bot.* **76** 657—665.

Chen J., Moinard M., Xu J., Wang S., Foulongne-Oriol M., Zhao R., Hyde K.D., Callac P. (2016) Genetic analyses of the internal transcribed spacer sequences suggest introgression and duplication in the medicinal mushroom *Agaricus subrufescens*. *PLOS ONE* **11** e0156250.

Chen L., VanBuren R., Paris M., Zhou H., Zhang X., Wai C.M., Yan H., et al. (2019) The *bracteatus* pineapple genome and domestication of clonally propagated crops. *Nat. Genet.* **51** 1549—1558.

Chudoba J., Křenková I., Mulač M., Ruda M., Sitera J. (2017) The Czech National Grid Infrastructure. *J. Phys.: Conf. Ser.* **898** 082042.

Dodsworth S., Chase M.W., Kelly L.J., Leitch I.J., Macas J., Novák P., Piednoël M., Weiss-Schneeweiss H., Leitch A.R. (2015) Genomic repeat abundances contain phylogenetic signal. Syst. Biol. **64** 112—126.

Emadzade K., Jang T.S., Macas J., Kovařík A., Novák P., Parker J., Weiss-Schneeweiss H. (2014) Differential amplification of satellite *PaB6* in chromosomally hypervariable *Prospero autumnale* complex (Hyacinthaceae) *Ann. Bot.* **114** 1597-1608

Favoreto F.C., Carvalho C.R., Lima A.B.P., Ferreira A., Clarindo W.R. (2012) Genome size and base composition of Bromeliaceae species assessed by flow cytometry. *Plant Syst. Evol.* **298** 1185—1193.

Fleischmann A., Michael T.P., Rivadavia F., Sousa A., Wang W., Temsch E.M., Greilhuber J., Müller K.F., Heubl G. (2014) Evolution of genome size and chromosome number in the carnivorous plant genus *Genlisea* (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms. *Ann. Bot.* **114** 1651—1663.

Gao D., Fox-Fogle E. (2023) Identification of transcriptionally active transposons in barley. *BMC Genom. Data* **24** 64.

Grandbastien M.-A., Spielmann A., Caboche M. (1989) Tnt1, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics. *Nature* **337** 376—380.

Gitai J., Paule J., Zizka G., Schulte K., Benko-Issepon A.M. (2014) Chromosome numbers and DNA content in Bromeliaceae: additional data and critical review. *Bot. J. Linn. Soc.* **176** 349—368.

Givnish T.J., Barfuss M.H.J., van Ee B., Riina R., Schulte K., Horres R., Gonsiska P.A., et al. (2011) Phylogeny, adaptive radiation, and historical biogeography in Bromeliaceae: insights from an eight-locus plastid phylogeny. *Am. J. Bot.* **98** 872—895.

Goffová I., Fajkus J. (2021) The rDNA loci — intersections of replication, transcription, and repair pathways. *Int. J. Mol. Sci.* **22** 1302.

Groot Crego C., Hess J., Yardeni G., de la Harpe M., Priemer C., Beclin F., Saadain S., et al. (2024, in press) CAM evolution is associated with gene family expansion in an explosive bromeliad radiation. *Plant Cell*

Haberer G., Kamal N., Bauer E., Gundlach H., Fischer I., Seidel M.A., Spannagl M., et al. (2020) European maize genomes highlight intraspecies variation in repeat and gene content. *Nat. Genet.* **52** 950—957.

de la Harpe M., Paris M., Hess J., Barfuss M.H.J., Serrano-Serrano M.L., Ghatak A., Chaturvedi P., et al. (2020) Genomic footprints of repeated evolution of CAM photosynthesis in a Neotropical species radiation. *Plant Cell Environ.* **43** 2987—3001.

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, New York

Hu X., Yu F., Huang Y., Sun L., Li X., Yang S., Chen K., et al. (2019) Characterization analysis of the 35S rDNA intergenic spacers in *Erianthus arundinaceus*. *Gene* **694** 63—70.

Huang X., Madan A. (1999) CAP3: A DNA sequence assembly program. *Genome Res.* **9** 868—877.

Langmead B., Salzberg S.L. (2012) Fast gapped-read alignment with Bowtie2. *Nat. Methods* **9** 357—359.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25** 2078—2079.

Li X., Waterman M.S. (2003) Estimating the Repeat Structure and Length of DNA Sequences using *l*-tuples. *Genome Res.* **13** 1916—1922.

Luebert F., Weigend M. (2014) Phylogenetic insights into Andean plant diversification. *Front. Ecol. Evol.* **2.**

Luo X., Chen S., Zhang Y. (2022) PlantRep: a database of plant repetitive elements. *Plant Cell Rep.* **41** 1163—1166.

Marçais G., Kingsford C. (2011) A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27** 764—770.

Mayer K.F.X., Waugh R., Langridge P., Close T.J., Wise R.P., Graner A., Matsumoto T., et al. (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature* **491** 711—717.

McClintock B. (1950) The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. U.S.A.*, June 1950, **36** 344—355.

Ming R., Van Buren R., Wai C.M., Tang H., Schatz M.C., Bowers J.E., Lyons E., et al. (2015) The pineapple genome and the evolution of CAM photosynthesis. *Nat. Genet.* **47** 1435—1442.

Müller L.B., Zotz G., Albach D.C. (2019) Bromeliaceae subfamilies show divergent trends of genome size evolution. *Sci. Rep.* **9** 5136.

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. *Mob. DNA* **10** 1.

Novák P., Neumann P., Macas J. (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinform.* **11** 378.

Novák P., Neumann P., Pech J., Steinhaisl J., Macas J. (2013) RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next generation sequence reads. *Bioinformatics* **29** 792—793.

Novák P., Robledillo L.A., Koblizková A., Vrbová I., Neumann P., Macas J. (2017) TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Res.* **45** e111.

Novák P., Neumann P., Macas J. (2020) Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. *Nat. Protoc.* **15** 3745—3776.

Nowoshilow S., Schloissnig S., Fei J.F., Dahl A., Pang A.W.C., Pippel M., Winkler S., et al. (2018) The axolotl genome and the evolution of key tissue formation regulators. *Nature* **554.**

Nurk S., Koren S., Rhie A., Rautiainen M., Bzikadze A.V., Mikheenko A., Vollger M.R., et al. (2022) The complete sequence of a human genome. *Science* **376** 44—53.

Pellicer J., Fay M.F., Leitch I.J. (2010) The largest eukaryotic genome of them all?. *Bot. J. Linn. Soc.* **164** 10—15.

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, Genet.* **10** 3047—3060.

Quesneville H. (2020) Twenty years of transposable element analysis in the *Arabidopsis thaliana* genome. *Mob. DNA* **11** 28.

Quinlan A.R., Hall I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26** 841—842.

Rosselló J.A., Maravilla A.J., Rosato M. (2022) The nuclear 35S rDNA world in plant systematics and evolution: a primer of cautions and common misconceptions in cytogenetic studies. *Front. Plant Sci.* **13** 788911.

Šatović-Vukšić E., Plohl M. (2023) Satellite DNAs - from localized to highly dispersed genome components. *Genes* **14** 742.

Shapiro J.A., von Sternberg R. (2005) Why repetitive DNA is essential to genome function. *Biol. Rev. Camb. Philos. Soc.* **80** 227—250.

Sonnhammer E.L., Durbin R. (1996) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167** GC1-10.

Sun H., Ding J., Piednoël M., Schneeberger K. (2017) findGSE: estimating genome size variation within human and *Arabidopsis* using *k*-mer frequencies. *Bioinformatics* **34** 550—557.

Sun L., Jing Y., Liu X., Li Q., Xue Z., Cheng Z., Wang D., He H., Qian W. (2020) Heat stress-induced transposon activation correlates with 3D chromatin organization rearrangement in *Arabidopsis*. *Nat. Commun.* **11** article no. 1886.

Temsch E.M. (2003) Genome size variation in plants with special reference to the genus *Arachis* (Fabaceae), and the methodological improvement by application of a

new optical immersion-gel in flow cytometry. Ph.D. Thesis, University of Vienna, Vienna, Austria

Thomas J., Pritham E.J. (2015) Helitrons, the eukaryotic rolling-circle transposable elements. *Microbiol. Spectr.* **3**.

Till W. (1992) Systematics and evolution of the tropical-subtropical *Tillandsia* subgenus *Diaphoranthema* (Bromeliaceae). *Selbyana* **13** 88—94.

Wegrzyn J.L., Liechty J.D., Stevens K.A., Wu L.S., Loopstra C.A., Yasquez-Gross H.A., Dougherty W.M., et al. (2014) Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. *Genetics* **196** 891—909.

Wicker T., Sabot F., Hua-Van A., Bennetzen J.L., Capy P., Chalhoub B., Flavell A., et al. (2007) A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8** 973—982.

6. Supplementary Material

Table S1. Composition of repetitive DNA fractions expressed as absolute sizes (in Mb).

		T. australis	*T. fasciculata*	*T. ionantha*	*T. leiboldiana*	*T. propagulifera*
Genome size		935.4	797.7	736.5	1,111.8	885.0
Mobile elements - Class I						
LTR/Ty1-*copia*	SIRE	175.5	131.7	105.2	236.4	84.5
	other copia	21.6	19.6	18.7	19.0	19.4
LTR/Ty3-gypsy	Tekay	367.3	295.6	279.1	527.0	448.9
	other gypsy	11.6	2.4	1.9	1.3	4.1
LTR/Cassandra-TRIM		2.4	2.5	2.4	5.7	1.7
LTR/unclassified		10.9	5.1	4.6	6.5	12.2
Non-LTR	LINE	—	0.4	0.4	0.3	—
Mobile elements - Class II						
Subclass I	TIR/ MITE	12.5	5.8	4.5	2.2	6.1
	other TIR	9.9	8.9	9.5	14.3	9.6
Subclass II	Helitron	1.3	0.7	2.7	0.7	0.7
"mixed" clusters - Class I & Class II	Tork & hAT	1.4	1.7	1.8	2.2	1.6
Tandem repeats						
rDNA	35S	3.9	2.2	15.2	4.3	4.1
	5S	2.5	1.9	7.0	2.9	0.7
Satellite DNAs		0.5	1.3	1.1	—	0.3
Unclassified repeats		50.7	42.3	29.0	27.0	40.0

		T. australis	*T. fasciculata*	*T. ionantha*	*T. leiboldiana*	*T. propagulifera*
Total top clusters		672.0	522.2	483.1	849.6	633.9
Small clusters		122.0	124.4	126.2	100.6	113.5
Singletons		141.3	151.0	127.3	161.6	137.6

Abstract

The genus *Tillandsia* has long been studied as an example of an adaptive radiation, with its ~600 species showing a wide diversity of ecotypes, and the repeated independent evolution of CAM photosynthesis being of especial interest. Species in this genus also have a wide range of genome sizes. This project addresses the pattern of genome size evolution in selected species in the genus, and how this is linked to repeatome evolution. Genome sizes of 21 species (16 in *Tillandsia* and five in other Tillandsioideae genera) were measured using flow cytometry and then estimated by two methods, the exon alignment method and the *k*-mer method, both using short read DNA sequence data. Comparison of the estimated genome sizes with measured values allowed the conclusion that both estimation methods showed systematic errors. The overestimates were due to the too large phylogenetic distance between *Tillandsia* and the *Ananas* reference genome used in the exon method, and to the insufficient depth of coverage of the DNA sequence data used in the *k*-mer method. Further, the repeatomes of five *Tillandsia* species were analyzed using the RepeatExplorer pipeline. The most common class of repeats recovered in all species was the LTR-retrotransposon, in particular Ty3-*gypsy*/Tekay and Ty1-*copia*/SIRE. DNA transposons, Cassandra-TRIMs, MITEs, and satellite DNAs were also found, but in much smaller amounts. Comparative analysis of the five species revealed the presence of distinct lineages of retrotransposons, amplified to differing extent in the individual species, suggesting that proliferation of genomic repeats has continued independently after the divergence of these species from each other. The data clearly showed a direct correlation between repeat abundances and genome sizes, and it was also shown that repeat profiles are informative of the phylogeny of *Tillandsia*.

Zusammenfassung

Die Gattung *Tillandsia* wird seit langem als Beispiel für eine adaptive Radiation untersucht. Ihre etwa 600 Arten weisen eine große Vielfalt an Ökotypen auf, wobei die wiederholte unabhängige Evolution der CAM-Photosynthese von besonderem Interesse ist. Außerdem weisen Arten dieser Gattung eine große Bandbreite an Genomgrößen auf. Dieses Projekt befasst sich mit dem Muster der Genomgrößenevolution bei ausgewählten Arten der Gattung und dessen Zusammenhang mit der Repeatom-Evolution. Durchflusszytometrische Messung der Genomgrößen von 21 Arten (16 bei *Tillandsia* und fünf bei anderen Gattungen der Tillandsioideae) wurde gefolgt von zwei Schätzmethoden, der Exon-Alignment-Methode und der *k*-mer-Methode, die beide kurze DNA-Sequenzdaten verwenden. Ein Vergleich der geschätzten Genomgrößen mit den gemessenen Werten ließ auf einen systematischen Fehler bei beiden angewandten Schätzmethoden schliessen. Die Überschätzungen waren im Fall der Exon-Methode auf die zu große phylogenetische Distanz zwischen *Tillandsia* und dem *Ananas*-Referenzgenom zurückzuführen, beziehungsweise auf die unzureichende Abdeckungstiefe der DNA-Sequenzdaten, im Fall der *k*-mer-Methode. Darüber hinaus erfolgte eine Analyse der Repeatome von fünf *Tillandsia*-Arten mithilfe der RepeatExplorer-Pipeline. Die häufigste Klasse von Repeats, die in allen Arten gefunden wurde, war das LTR-Retrotransposon, insbesondere Ty3-*gypsy*/Tekay und Ty1-*copia*/SIRE. DNA-Transposonen, Cassandra-TRIMs, MITEs und Satelliten-DNAs wurden ebenfalls gefunden, allerdings in viel geringeren Mengen. Eine vergleichende Analyse der fünf Arten ergab das Vorhandensein unterschiedlicher Linien von Retrotransposonen, die in den einzelnen Arten in unterschiedlichem Ausmaß verstärkt wurden, was auf eine unabhängige Fortführung der Verbreitung genetischer Repeats nach der Divergenz dieser Art deutet. Die Daten zeigten eindeutig eine direkte Korrelation zwischen Repeat-Häufigkeit und Genomgröße. Zudem wurde gezeigt, dass Repeat-Profile Aufschluss über die Phylogenese von *Tillandsia* geben.